

# 확률

## 모듈 - 2

강사: 장순용 박사

광주인공지능사관학교 제 2기 (2021/06/16~2021/12/02) 용도로 제공되는 강의자료 입니다. 지은이의 허락 없이는 복제와 배포를 금합니다.

## 2. 확률 II:

### 2.1. 이산확률변수 & 확률분포.

2.2. 이산확률분포의 여러 종류.

2.3. 연속확률변수 & 확률밀도.

2.4. 연속확률밀도의 여러 종류.

2.5. 결합확률과 상관계수.

## 확률변수 : 정의

### 확률변수:

- 확률변수 (random variable): 확률실험의 각 결과에 실수를 부여하는 함수.
- 확률변수의 값은 하나의 수치로 나타낸 사건이다.  
**예).** 동전을 한번 던지는 실험에서 앞면(H)가 나오면 1 뒷면(T)이 나오면 0.

## 확률변수 : 유형

확률변수의 유형: 크게는 두가지의 유형이 있음.

- 이산확률변수 (discrete random variable): 셀 수 있는 가지수의 값을 가지는 확률변수.

예). 주사위를 던져서 나오는 눈의 수:  $\{1, 2, 3, 4, 5, 6\}$

- 연속확률변수 (continuous random variable): 셀 수 없는 (무한대) 가지수의 값을 가지는 확률변수.

예). 1년 연봉, 성인남성의 신장, 등.

### 확률분포의 유형:

- 이산확률분포함수 (discrete probability distribution): 이산확률변수가 가지는 값과 이것의 확률 사이의 대응 관계.

→ 보통 확률변수는 영문 **대문자**로 표기하고 확률변수의 값은 영문 **소문자**로 표기한다.

**예**). 확률변수  $X$ , 확률변수의 값  $x$ .

→ 확률변수  $X$ 의 값이  $x$ 일 확률은  $P(X = x)$  또는  $P(x)$ 와 같이 표기한다.

### 확률분포의 유형:

- 연속 확률 분포 함수 / 확률 밀도 함수 (continuous probability distribution/probability density function): 이것을 **사용하여** 연속 확률 변수의 값이 특정 **구간에 속할 확률**을 나타냄. (이후 단원에서 자세히 다루기로 함)

## 이산확률분포 : 필수조건

이산확률분포의 필수조건:

- $0 \leq P(x) \leq 1$
- $\sum_{all\ x_i} P(x_i) = 1$

예). 인공신경망의 출력층의 활성화 함수로 사용되는 Softmax.

### 누적확률:

- 확률변수  $X$ 의 개별값이 실현되는 확률을 확률분포  $P(x)$ 로 나타내었다.
- 누적확률은 확률변수의 값이 특정 구간에 속할 확률이다.

예).  $P(X \geq 80)$

$$P(X \leq 70)$$

$$P(X \text{는 } 50 \text{ 이상 } 80 \text{ 이하})$$

$$P(X \text{는 최소값 이상 최고값 이하}) = 1$$

- 누적확률은 0 이상 1이하의 수치이다.



## 모집단과 모수

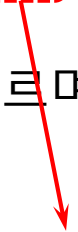
### 모집단과 모수:

- 모집단 (population): 분석 대상 전체 (현실적 또는 이상적).
- 모수 (parameter): 모집단의 특성을 의미함.  
예). 모평균, 모분산, 모표준편차, 등.
- 확률분포함수를 사용해서 모수를 계산할 수 있다.

## 이산확률변수 : 모평균

이산확률변수의 모평균  $\mu$ :

- 이산확률변수가 가질 수 있는 값들에 확률을 **가중치**로 곱해서 평균을 구한 것 ( $\cong$  더한 것).
- 확률변수  $X$ 의 **기대값** (expected value)이라고도 부르며  $E[X]$ 로 나타낸다.

$$\mu = E[X] = \sum_{all\ x} x P(x)$$


## 이산확률변수 : 모분산

이산확률변수의 모분산  $\sigma^2$  :

- 모평균을 기준으로한 **편차의 제곱**에 확률을 **가중치**로 곱해서 평균을 구한 것 ( $\approx$  더한 것).
- 확률변수  $X$ 의 모분산을  $Var(X)$ 와 같이 나타내기도 한다.

$$\sigma^2 = Var(X) = \sum_{all\ x} (x - \mu)^2 P(x)$$

- 모분산의 간편 수식:

$$\begin{aligned}\sigma^2 = Var(X) &= \left( \sum_{all\ x} x^2 P(x) \right) - \mu^2 \\ &= E[X^2] - (E[X])^2\end{aligned}$$

- 모표준편차:  $\sigma = \sqrt{\sigma^2}$

## 이산확률분포 : 예제 #0201

다음 표와 같이 성적이 분포해 있을 때, 모평균과 모분산을 계산 해 보세요:

점수 (X)	확률 $P(x)$
60	1/30
70	9/30
80	11/30
90	7/30
100	2/30
합계	1

- $$\begin{aligned}\mu &= 60 P(60) + 70 P(70) + 80 P(80) + 90 P(90) + 100 P(100) \\ &= 60 \times \frac{1}{30} + 70 \times \frac{9}{30} + 80 \times \frac{11}{30} + 90 \times \frac{7}{30} + 100 \times \frac{2}{30} = 80\end{aligned}$$

## 이산확률분포 : 예제 #0201

다음 표와 같이 성적이 분포해 있을 때, 모평균과 모분산을 계산 해 보세요:

점수 (X)	확률 $P(x)$
60	1/30
70	9/30
80	11/30
90	7/30
100	2/30
합계	1

- $\sigma^2 = \{60^2 P(60) + 70^2 P(70) + 80^2 P(80) + 90^2 P(90) + 100^2 P(100)\} - 80^2$   
 $= 3600 \times \frac{1}{30} + 4900 \times \frac{9}{30} + 6400 \times \frac{11}{30} + 8100 \times \frac{7}{30} + 10000 \times \frac{2}{30} - 6400 = 93.333$
- $\sigma = \sqrt{93.333} = 9.66$

## 2. 확률 II:

2.1. 이산확률변수 & 확률분포.

2.2. 이산확률분포의 여러 종류.

2.3. 연속확률변수 & 확률밀도.

2.4. 연속확률밀도의 여러 종류.

2.5. 결합확률과 상관계수.

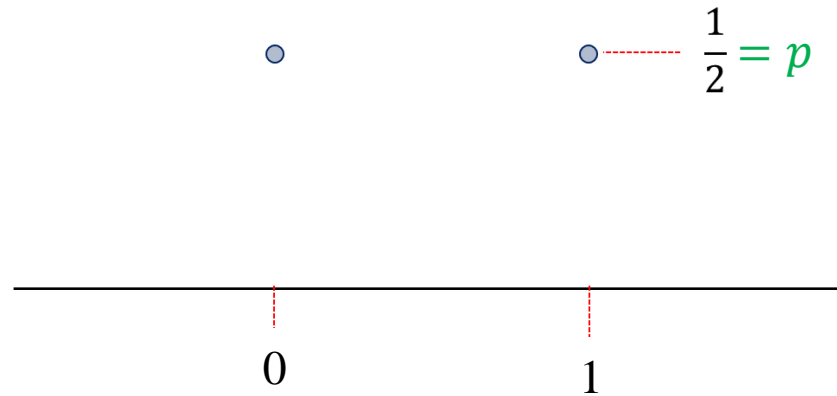
## 이산확률분포 : 베르누이

베르누이 확률분포 (Bernoulli):



## 이산확률분포 : 베르누이

베르누이 확률분포 (Bernoulli):



$$P(x) = p^x(1 - p)^{1-x}$$

평균 :  $p$

분산 :  $p(1 - p)$

표준편차 :  $\sqrt{p(1 - p)}$

베르누이 분포 (“1회 동전 던지기”)



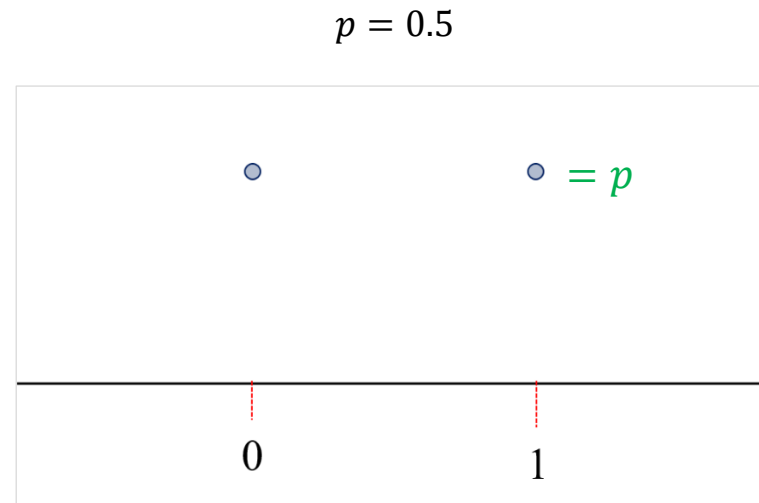
### 베르누이 확률분포 (Bernoulli):

- 베르누이 시행에는 두개의 가능한 값이 있다.  
예). 1 또는 0, 동전의 앞면(H) 또는 뒷면(T), “성공” 또는 “실패”.
- 확률변수  $X$ 가 베르누이 확률분포를 따른다는 것을 다음과 같이 표기할 수 있다.

$$X \sim Ber(p)$$

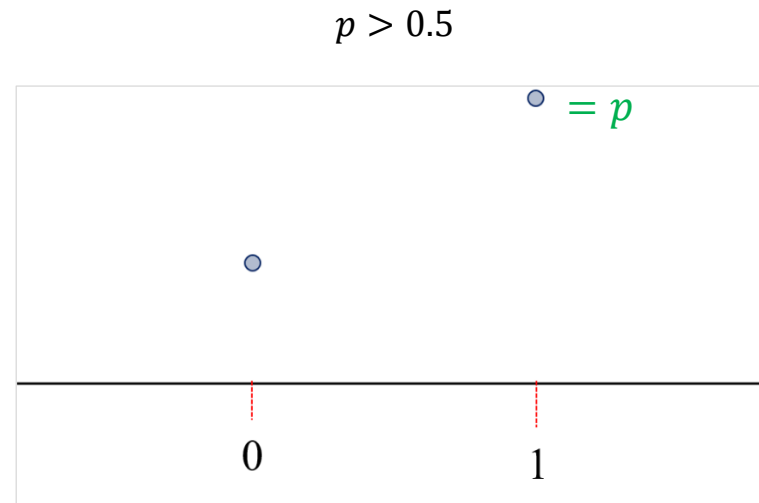
## 이산확률분포 : 베르누이

베르누이 확률분포 (Bernoulli):



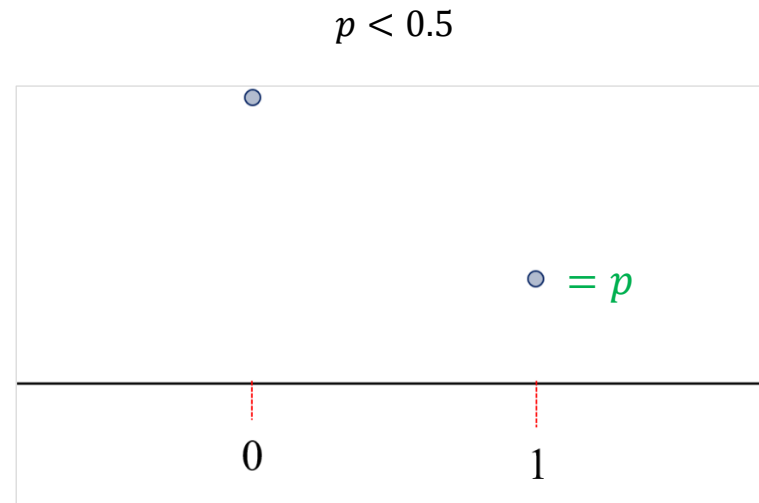
## 이산확률분포 : 베르누이

베르누이 확률분포 (Bernoulli):

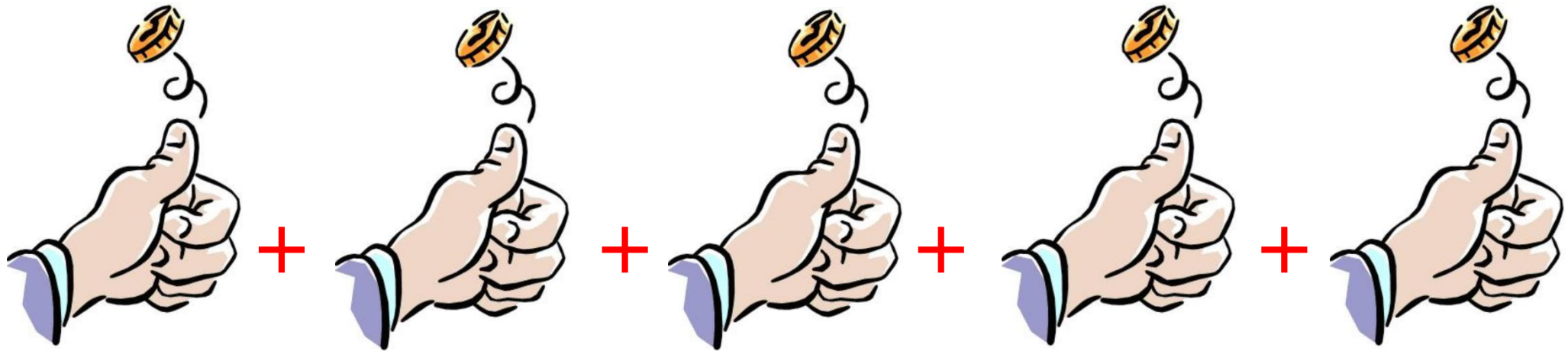


## 이산확률분포 : 베르누이

베르누이 확률분포 (Bernoulli):

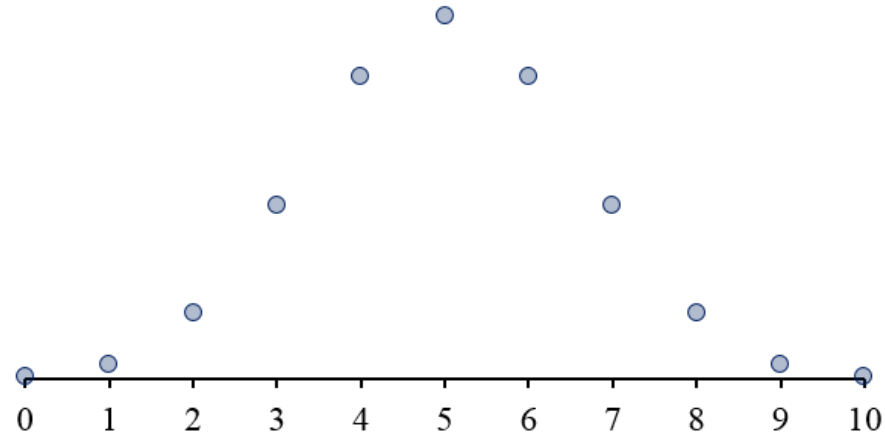


이항확률분포 (Binomial):



## 이산확률분포 : 이항

이항확률분포 (Binomial):



이항확률분포 (“ $n$ 회 동전 던지기”)

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\text{평균} : np$$

$$\text{분산} : np(1-p)$$

$$\text{표준편차} : \sqrt{np(1-p)}$$

## 이항확률분포 (Binomial):

- 이항확률변수  $X_{bin}$ 은 0 또는 1의 값을 갖는  $n$ 개의 베르누이확률변수  $X_{Ber}$ 를 더한 것.

$$X_{bin} = X_{Ber} + X_{Ber} + \cdots + X_{Ber}$$

←  $n$  개 →

예). 동전 하나를  $n$  번 던져서 앞면(H)이 나온 횟수를 집계한 것.

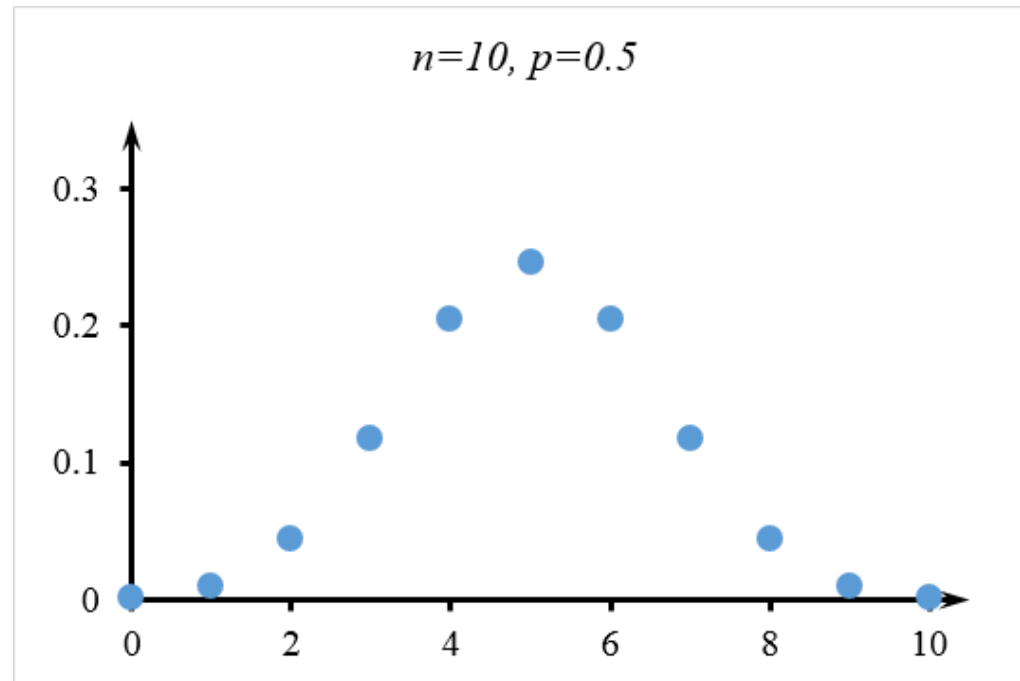
예).  $n$ 회 시행하여 “성공”한 횟수를 더한 것.

- 개개의 베르누이확률변수는 독립적이다.  $\Rightarrow$  동전 던지기는 이전 결과와는 무관하다.
- 확률변수  $X$ 가 이항확률분포를 따른다는 것을 다음과 같이 표기할 수 있다.

$$X \sim \text{Bin}(n, p)$$

## 이산확률분포 : 이항

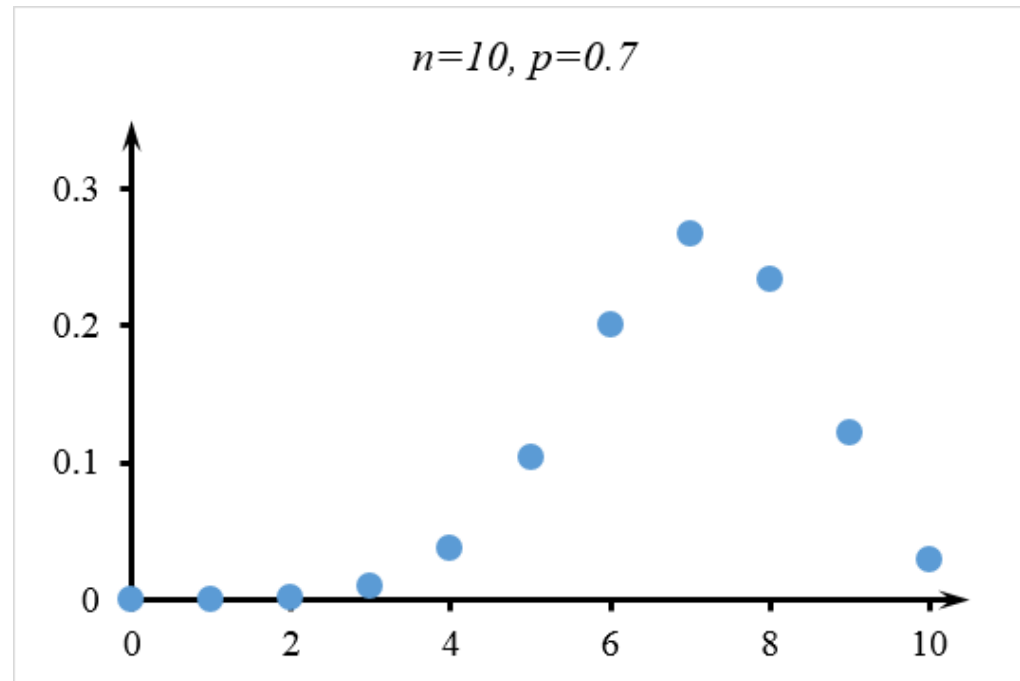
이항확률분포 (Binomial): 좌우 대칭인 경우.





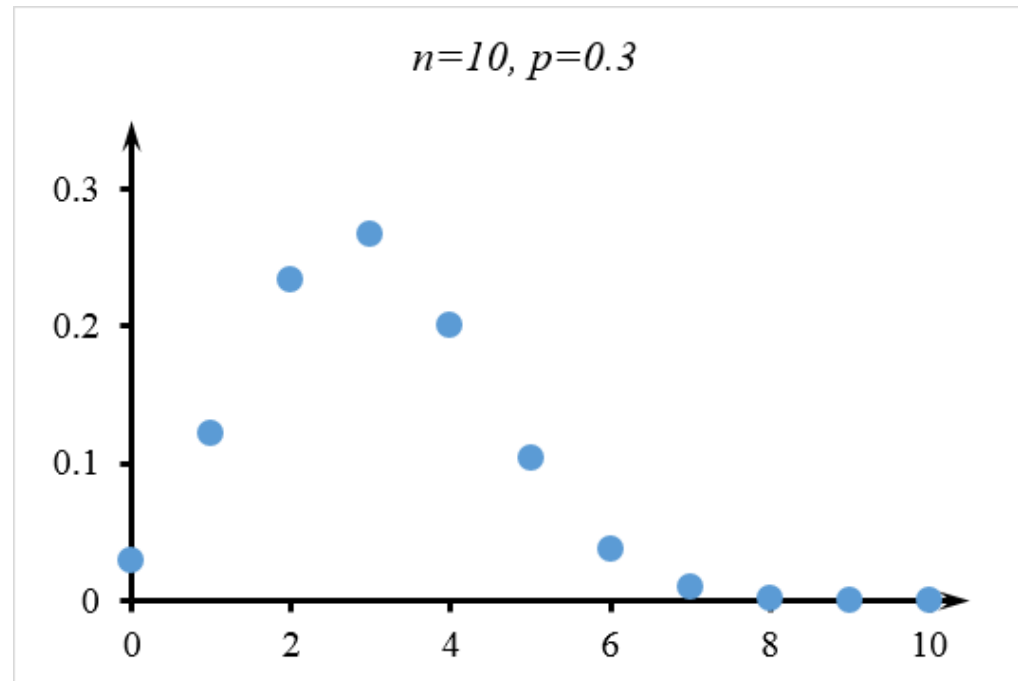
## 이산확률분포 : 이항

이항확률분포 (Binomial): 확률이 오른쪽으로 쏠리는 경우.



## 이산확률분포 : 이항

이항확률분포 (Binomial): 확률이 왼쪽으로 쏠리는 경우.



난제를 풀어봅시다.



난제 = *brainteaser*

난제를 풀어봅시다.

**질문:** 다음과 같은 룰의 동전 던지기 게임에 참가하겠습니까?

**동전 던지기 게임의 룰:**

- 앞면이 나오면 수익 100\$ 이고, 반대로 뒷면이 나오면 수익 0\$이다.
- 매번 게임에 참가하는 비용은 40\$이다.

## 이산확률의 활용

난제를 풀어봅시다.

그런데, 동전은 앞면이 나올 확률과 뒷면이 나올 확률이 같습니다.  
1회 기대수익은 다음과 같습니다.



$$\text{기대수익} = \frac{1}{2} \times 100\$ + \frac{1}{2} \times 0\$ = 50\$$$



기대수익 50\$가 비용 40\$보다 크니까 물론 게임에 참가?

## 이산확률의 활용

난제를 풀어봅시다.

잠깐만, 1회 수익의 리스크(변동성, 표준편차)를 계산해봅시다.



$$\text{리스크} = \sqrt{\frac{1}{2} \times (100 - 50)^2 + \frac{1}{2} \times (0 - 50)^2} = 50\$$$

## 이산확률의 활용

난제를 풀어봅시다.

샤프지수는 리스크 대비 초과수익을 나타냅니다.



$$\begin{aligned}\text{샤프지수} &= \frac{\text{초과수익}}{\text{리스크}} \\ &= \frac{(\text{기대수익} - \text{참가비용})}{\text{리스크}} \\ &= \frac{(50 - 40)}{50} = 0.2\end{aligned}$$



1 미만의 샤프지수는 불만족스럽습니다.

# 이산확률의 활용

난제를 풀어봅시다.

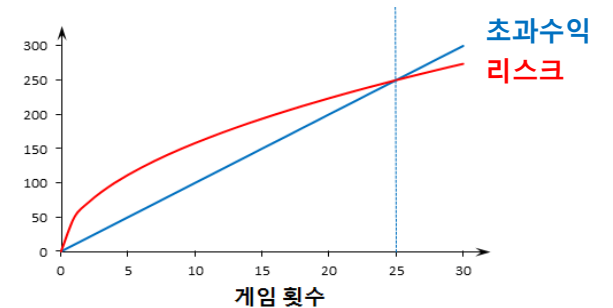
*그런데, 1 회 이상  $n$  회 동전 던지기를 할 수 있다면?*

- 초과수익은  $n$ 에 비례해서 증가합니다.
- 리스크(표준편차, 변동성)은  $\sqrt{n}$ 에 비례해서 증가합니다.
- 이항 분포 (binomial distribution)의 특성입니다.

↓

$$\text{샤프지수} = \frac{(50 - 40) \times n}{50 \times \sqrt{n}} = 0.2 \times \sqrt{n}$$

↓



25 회 이상 플레이하면 샤프지수는 1을 초과하게 됩니다. → **만족 !**



난제를 풀어봅시다: 확률 변수를 명시하여 다시한번 풀어본다.

- 다음과 같이  $p = 1/2$ 인 확률변수  $X_{Ber}$ 을 사용하여 **1 회** 동전 던지기의 “수익” 확률 변수  $X$ 를

모델링 해본다:  $X = 100 \times X_{Ber}$

$$\Rightarrow \text{기대수익 } \mu = E[X] = E[100 \times X_{Ber}] = 100 \times E[X_{Ber}] = 100 \times \frac{1}{2} = 50$$

$$\Rightarrow \sigma^2 = \text{Var}(X) = \text{Var}(100 \times X_{Ber}) = 10000 \times \text{Var}(X_{Ber}) = 10000 \times \frac{1}{2} \times \left(1 - \frac{1}{2}\right) = 2500$$

$$\Rightarrow \sigma = \sqrt{\sigma^2} = 50$$

$$\Rightarrow \text{샤프지수} = \frac{\text{초과수익}}{\text{리스크}} = \frac{(\text{기대수익} - \text{참가비용})}{\text{리스크}} = \frac{(50 - 40)}{50} = 0.2$$

## 이산확률의 활용

난제를 풀어봅시다: 확률 변수를 명시하여 다시한번 풀어본다.

- 다음과 같이  $p = 1/2$ 인 확률변수  $X_{bin}$ 을 사용하여  $n$  회 동전 던지기의 “수익” 확률 변수  $X$ 를

모델링 해본다:  $X = 100 \times X_{bin}$

$$\Rightarrow \text{기대수익 } \mu = E[X] = E[100 \times X_{bin}] = 100 \times E[X_{bin}] = 100 \times n \times \frac{1}{2} = 50 \times n$$

$$\Rightarrow \sigma^2 = \text{Var}(X) = \text{Var}(100 \times X_{bin}) = 10000 \times \text{Var}(X_{bin}) = 10000 \times n \times \frac{1}{2} \times \left(1 - \frac{1}{2}\right)$$

$$= 2500 \times n$$

$$\Rightarrow \sigma = \sqrt{\sigma^2} = 50 \times \sqrt{n}$$

$$\Rightarrow \text{샤프지수} = \frac{\text{초과수익}}{\text{리스크}} = \frac{(\text{기대수익} - \text{참가비용})}{\text{리스크}} = \frac{(50 \times n - 40 \times n)}{50 \times \sqrt{n}} = 0.2 \times \sqrt{n}$$

## 이산확률분포 : Python 함수

Python의 이산확률 함수:

명칭	함수
이항 (Binomial)	<code>scipy.stats.<b>binom</b>.pmf()</code> ← 확률분포 <code>scipy.stats.<b>binom</b>.cdf()</code> ← 누적확률 <code>scipy.stats.<b>binom</b>.ppf()</code> ← 분위수
푸아송 (Poisson)	<code>scipy.stats.<b>poisson</b>.pmf()</code> <code>scipy.stats.<b>poisson</b>.cdf()</code> <code>scipy.stats.<b>poisson</b>.ppf()</code>

## 실습 #0201

→ 이산확률분포에 대해서 알아봅니다. ←

→ 사용: **ex\_0201.ipynb** ←

## 2. 확률 II:

2.1. 이산확률변수 & 확률분포.

2.2. 이산확률분포의 여러 종류.

2.3. 연속확률변수 & 확률밀도.

2.4. 연속확률밀도의 여러 종류.

2.5. 결합확률과 상관계수.

## 연속확률변수

### 연속확률변수:

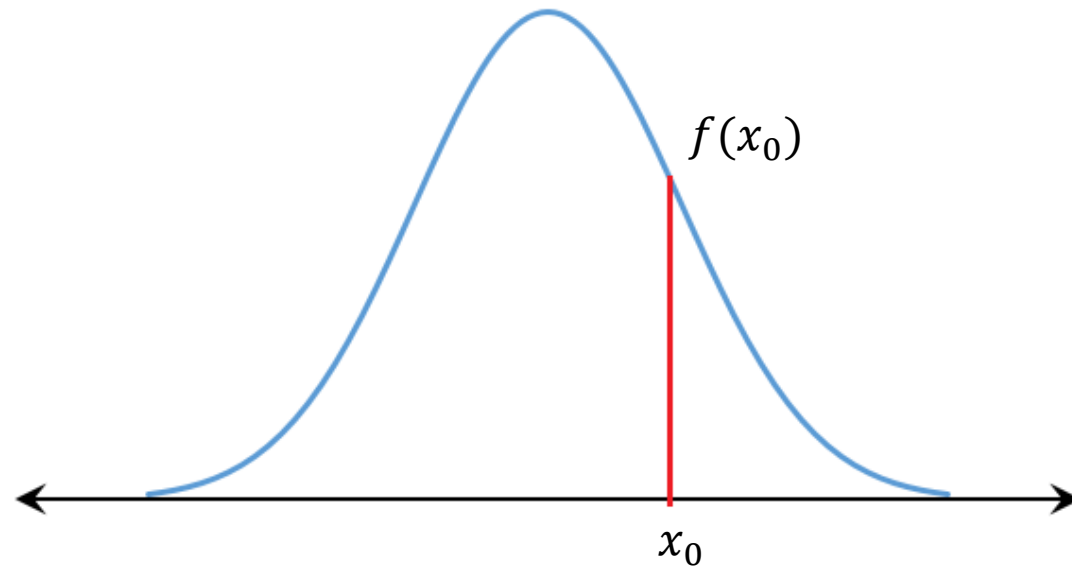
- 연속확률변수 (continuous random variable): 셀 수 없는 (무한대) 가지수의 값을 가지는 확률변수.  
**예).** 1년 연봉, 성인남성의 신장, 등.
- 연속확률변수의 경우 확률은 **실수 구간**에 대해서 정의되어 있음. 즉  $P(X = x_0)$ 와 같이 특정 위치에 대한 확률은 의미가 없고,  $P(x_1 \leq X \leq x_2)$ 와 같이  $X$ 가 어느 실수 구간에 있을 확률이 의미가 있다.

### 연속확률분포:

- 연속확률분포함수/확률밀도함수 (continuous probability distribution/probability density function):
  - 이산확률분포함수와는 다르게 이것 자체만으로는 확률의 **의미가 없다**.
  - 이것을 **사용하여** 연속확률변수의 값이 특정 구간에 속할 확률을 나타낼 수 있다.
  - 연속확률분포함수 또는 확률밀도함수를  $f(x)$ 와 같이 표기하여 실제 확률  $P(x)$ 와는 구분 짓도록 한다.

## 연속확률분포

연속확률분포:

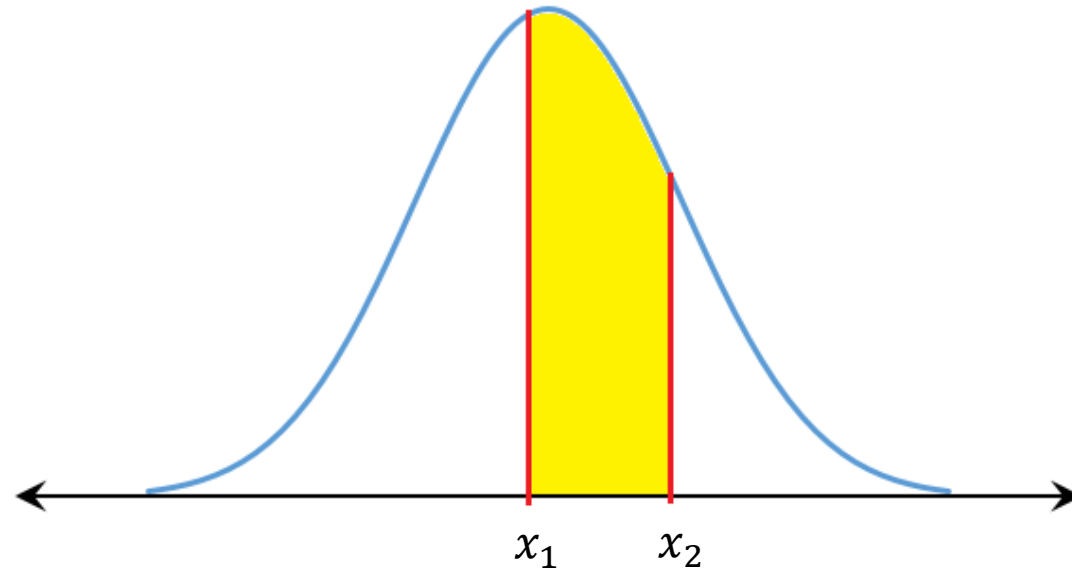


- $f(x_0)$ 에는 확률의 의미가 없다.



## 연속확률분포

연속확률분포:

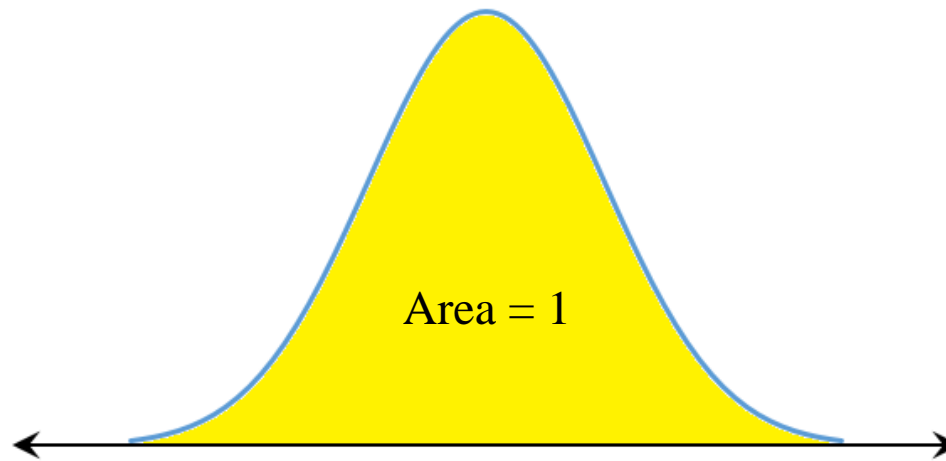


- $P(x_1 \leq X \leq x_2)$ 와 같이  $X$ 가 어느 실수 **구간**에 있을 확률이 의미가 있다 (음영).

## 연속확률분포 : 필수조건

연속확률분포의 필수조건:

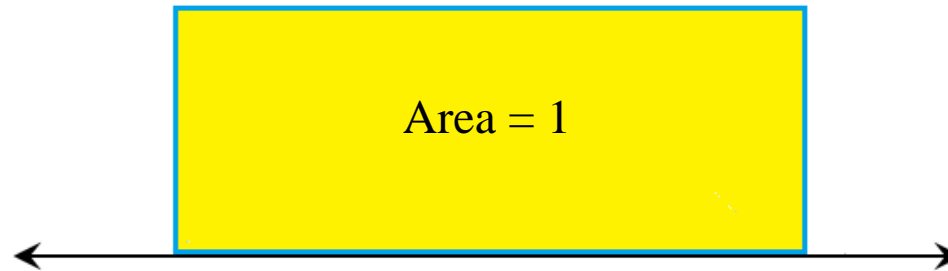
- $0 \leq f(x)$
- $f(x)$ 가 정의되어 있는 구간에서  $f(x)$  아래의 총 면적은 1과 같아야 한다.



## 연속확률분포 : 필수조건

연속확률분포의 필수조건:

- $0 \leq f(x)$
- $f(x)$ 가 정의되어 있는 구간에서  $f(x)$  아래의 총 면적은 1과 같아야 한다.



## 2. 확률 II:

2.1. 이산확률변수 & 확률분포.

2.2. 이산확률분포의 여러 종류.

2.3. 연속확률변수 & 확률밀도.

2.4. 연속확률밀도의 여러 종류.

2.5. 결합확률과 상관계수.

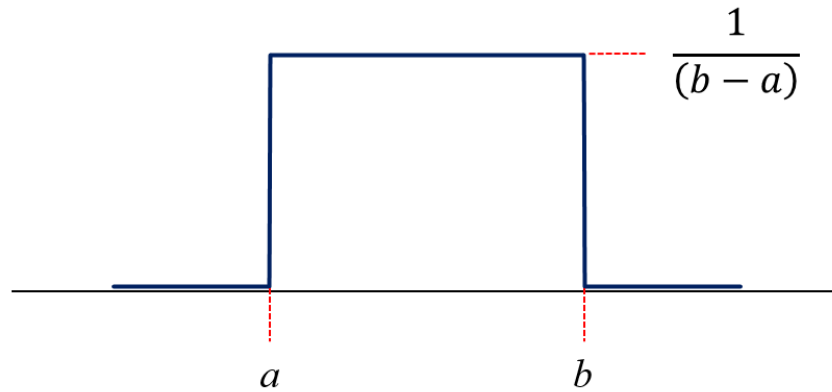
## 연속확률분포 : 유형별 용도

연속확률분포함수 유형별 용도 정리:

명칭	활용
정규분포	대표본 구간 추정. 대표본 평균 추론 (가설검정).
스튜던트 t 분포	소표본 구간 추정. 소표본 평균 추론 (가설검정). 선형회귀 계수 추론 (가설검정).
카이제곱 분포	분산 추론 (가설 검정). 범주형 자료를 정리한 도수표 추론 (가설검정). 범주형 자료를 정리한 분할표 추론 (가설검정).
F 분포	분산의 차이 비교 추론 (가설검정). 다수의 집단의 평균 비교 추론 (ANOVA). 선형회귀식의 설명력 추론 (가설검정)

## 연속확률분포 : 연속균등분포

연속균등확률분포함수 (Uniform):



$$f(x) = \frac{1}{(b-a)}$$

$$\text{평균} : \frac{1}{2}(a+b)$$

$$\text{분산} : \frac{1}{12}(b-a)^2$$

$$\text{표준편차} : \frac{1}{\sqrt{12}}(b-a)$$

**연속균등분포**

연속균등확률분포함수 (Uniform):

- 연속균등확률분포함수는 구간  $[a, b]$ 에 대해서 정의되어 있다:

$$f(x) = \frac{1}{(b - a)}$$

→ 이외의 구간에서는  $f(x) = 0$ 이다.

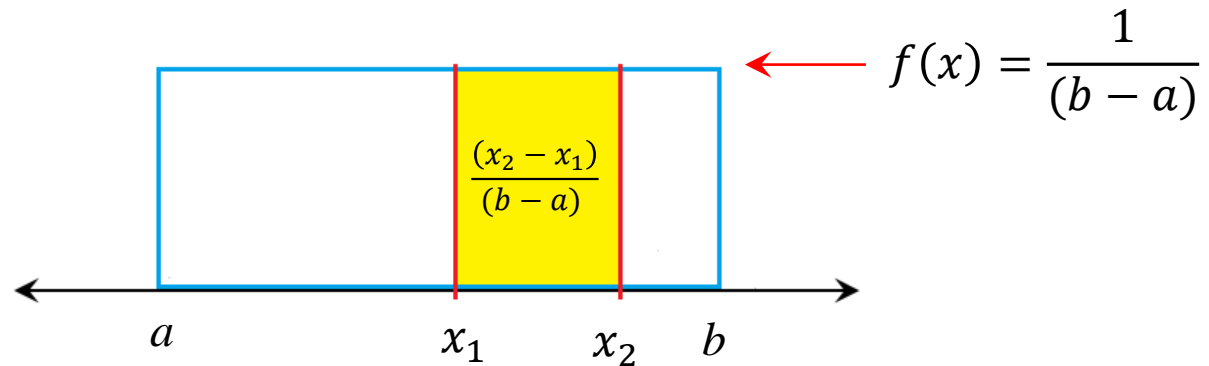
- 확률변수  $X$ 가 연속균등확률분포를 따른다는 것을 다음과 같이 표기할 수 있다.

$$X \sim Unif(a, b)$$

## 연속확률분포 : 연속균등분포

연속균등확률분포함수 (Uniform):

- 확률밀도가 균등하므로 확률은 주어진 구간의 폭  $(x_1 - x_2)$ 에 비례한다.

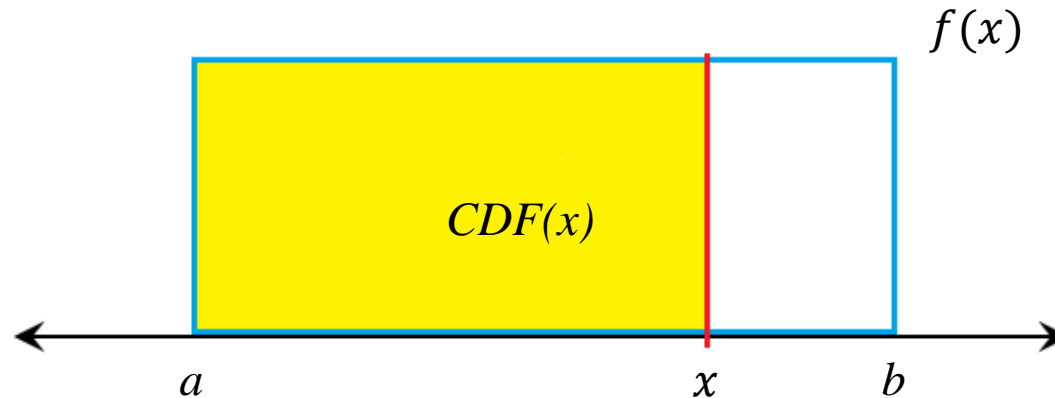




## 연속균등분포 : 누적확률함수

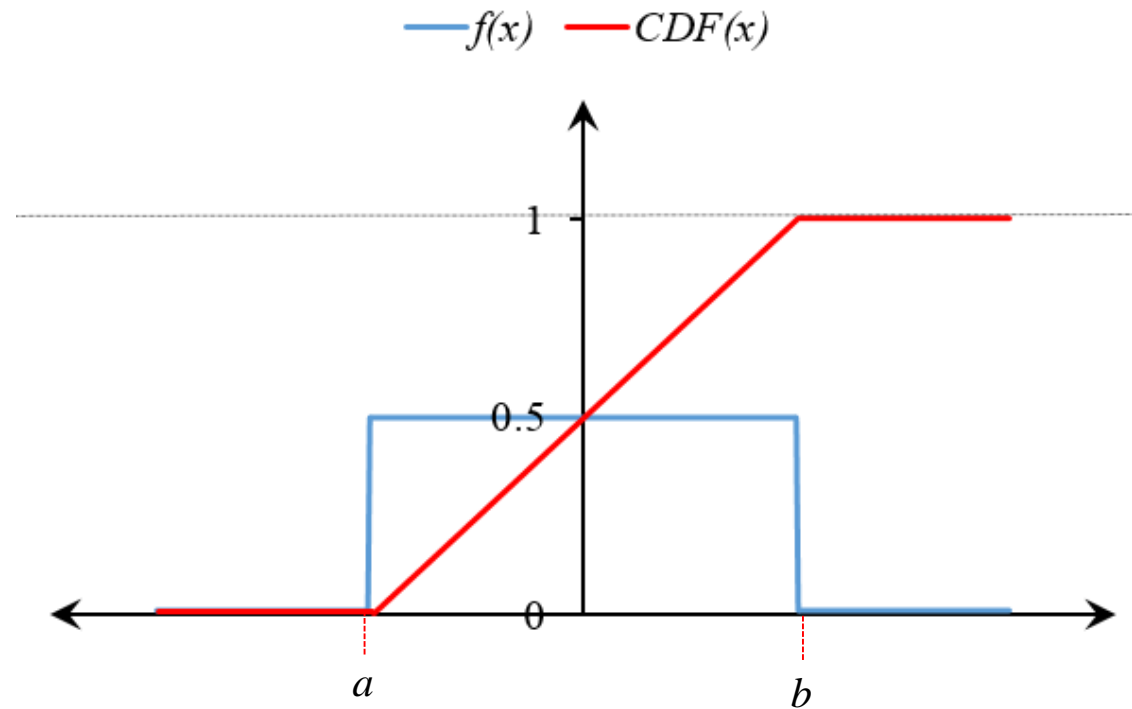
연속균등분포의 누적확률함수 (Cumulative Density Function, CDF):

- 연속균등분포의 누적확률  $CDF(x)$ 는 구간  $[a, x]$  에서  $f(x)$  아래의 면적 (■)과 같다.
- $CDF(x) = P(X \leq x)$ 이다.



## 연속균등분포 : 누적확률함수

연속균등분포의 누적확률함수 (Cumulative Density Function, CDF):



$CDF(x)$  는  $x$ 가 증가하면 1로 수렴한다.



백열전구의 수명 ( $X$ )은 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

1). 어느 백열전구가 사용시간 6000시간과 7000시간 사이에서 타버릴 확률은?

→ 확률밀도 함수는  $f(x) = \frac{1}{(7000-5000)} = 0.0005$ 이다. 그러므로

$$P(6000 \leq X \leq 7000) = (7000 - 6000) \times f(x) = 1000 \times 0.0005 = 0.5$$

2). 어느 백열전구의 수명이 5500시간 이하일 확률은?

$$\rightarrow P(X \leq 5500) = P(5000 \leq X \leq 5500) = (5500 - 5000) \times f(x) = 500 \times 0.0005 = 0.25$$



백열전구의 수명 ( $X$ )은 연속균등분포를 따르며 5000시간에서 7000시간 사이라고 한다. 다음 물음에 답하시오.

3). 어느 백열전구의 수명이 최소 사용시간 5500시간 이상일 확률은?

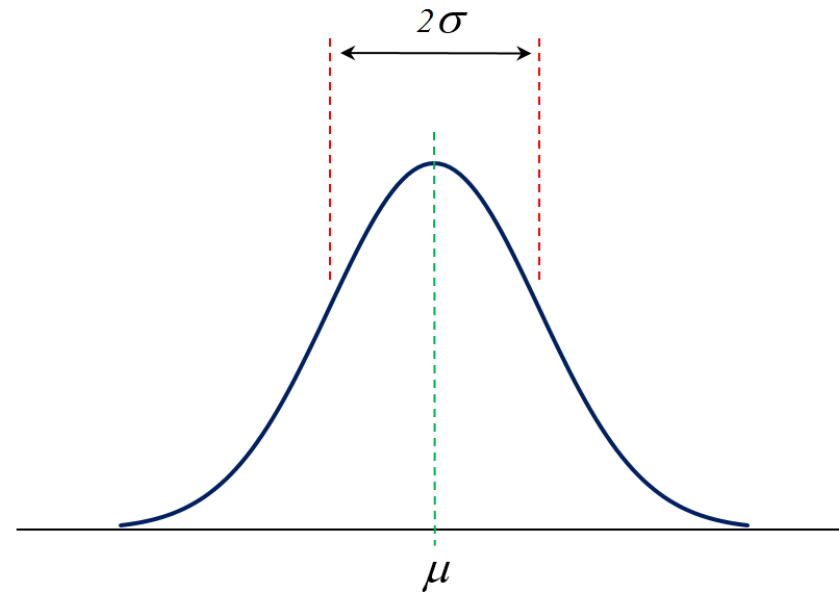
$$\rightarrow P(5500 \leq X) = P(5500 \leq X \leq 7000) = (7000 - 5500) \times f(x) = 1500 \times 0.0005 = 0.75$$

4). 어느 백열전구의 수명이 정확하게 6000시간일 확률은?

$$\rightarrow P(X = 6000) = P(\mathbf{6000} \leq X \leq \mathbf{6000}) = (6000 - 6000) \times f(x) = 0 \times 0.0005 = \mathbf{0}$$

## 연속확률분포 : 정규분포

정규확률분포함수 (Normal):



정규분포

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

평균 :  $\mu$

분산 :  $\sigma^2$

표준편차 :  $\sigma$

### 정규확률분포함수 (Normal):

- 정규확률분포함수는 구간  $(-\infty, +\infty)$ 에 대해서 정의되어 있다:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$e = 2.71828$$

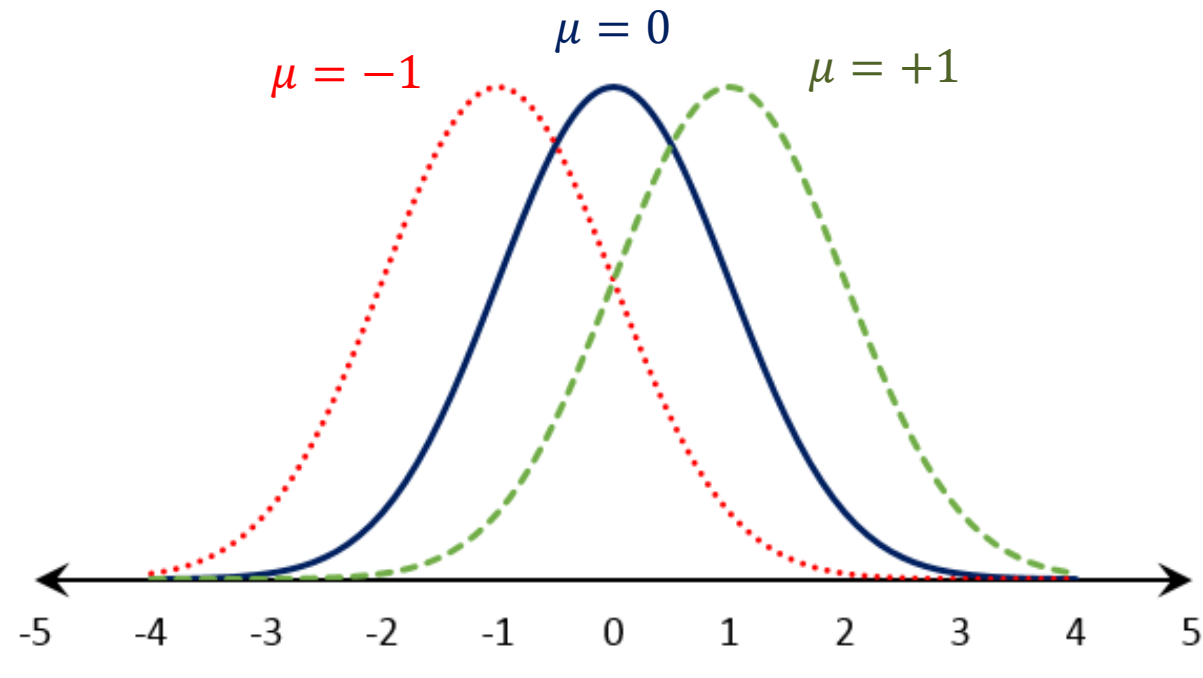
$$\pi = 3.141592$$

- 확률변수  $X$ 가 정규확률분포를 따른다는 것을 다음과 같이 표기할 수 있다.

$$X \sim N(\mu, \sigma^2)$$

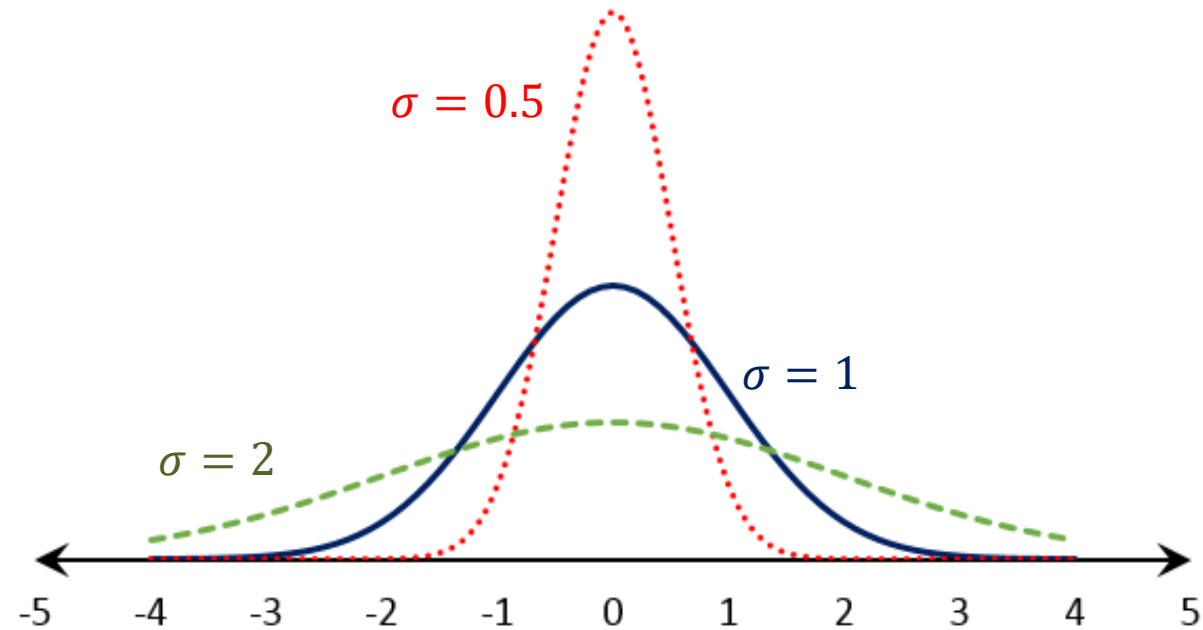
## 연속확률분포 : 정규분포

정규확률분포함수 (Normal):  $\mu$ 의 역할



## 연속확률분포 : 정규분포

정규확률분포함수 (Normal):  $\sigma$ 의 역할





### 표준정규확률분포함수 (Standard Normal):

- $\mu = 0$ 이고  $\sigma^2 = 1$ 인 정규확률분포를 **표준정규분포**라 한다:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$$e = 2.71828$$

$$\pi = 3.141592$$

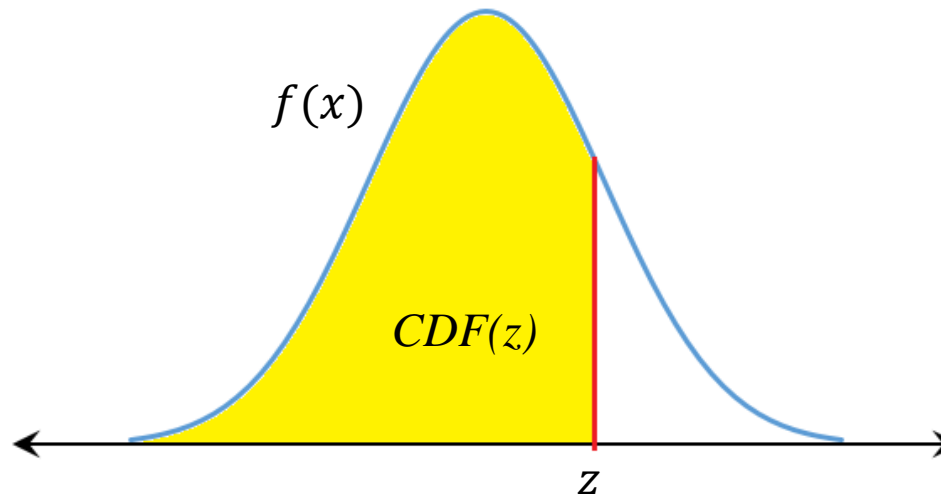
- 확률변수  $Z$ 가 **표준정규확률분포**를 따른다는 것을 다음과 같이 표기할 수 있다.

$$Z \sim N(0,1)$$

## 정규분포 : 누적확률함수

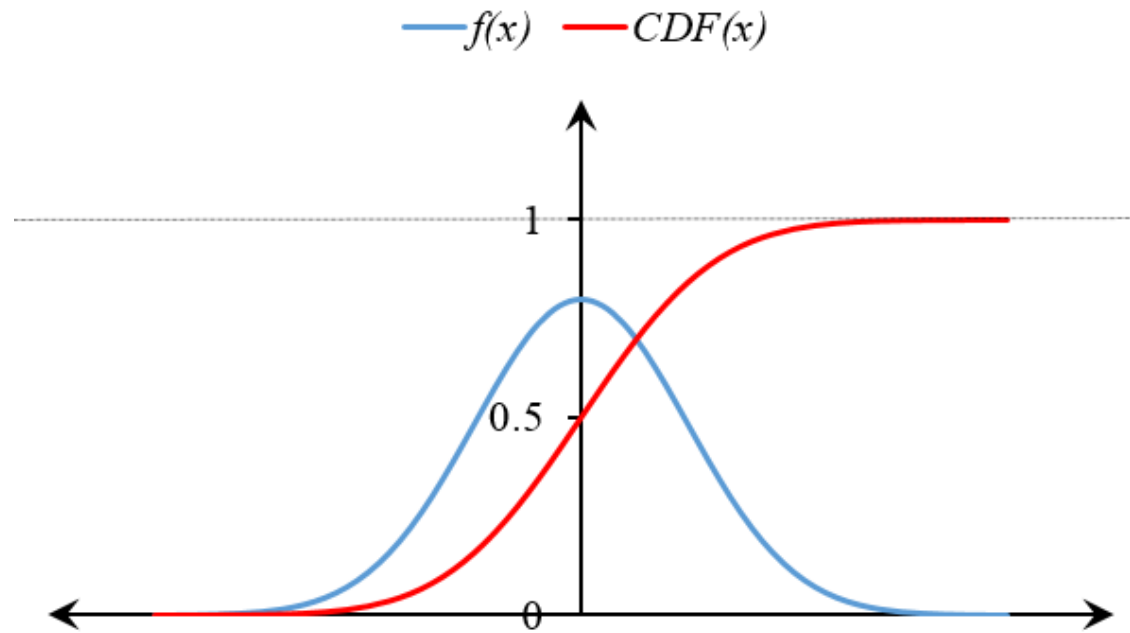
표준정규분포의 누적확률함수 (Cumulative Density Function, CDF):

- 표준정규분포의 누적확률  $CDF(z)$ 는 구간  $(-\infty, z]$  에서  $f(x)$  아래의 면적 (■)과 같다.
- $CDF(z) = P(Z \leq z)$ 이다.



## 정규분포 : 누적확률함수

표준정규분포의 누적확률함수 (Cumulative Density Function, CDF):

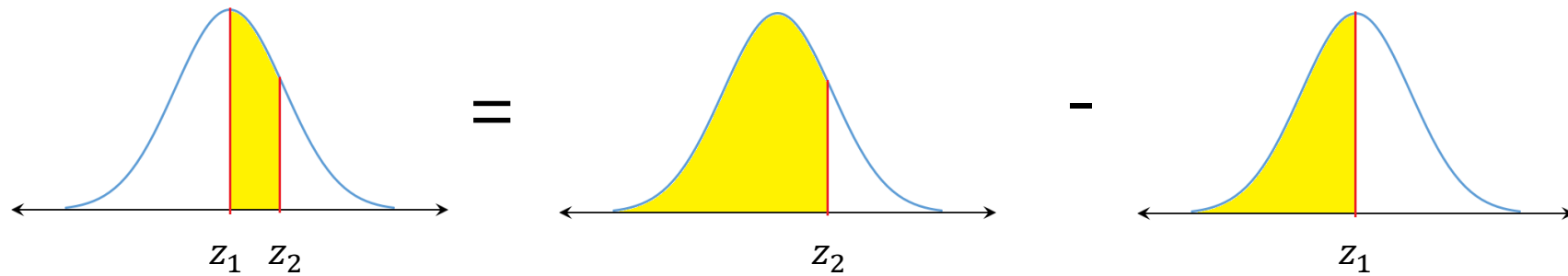


$CDF(x)$  는  $x$ 가 증가하면 1로 수렴한다.

## 정규분포 : 누적확률함수

표준정규분포의 누적확률함수 (Cumulative Density Function, CDF):

- $P(z_1 \leq Z \leq z_2)$ 와 같이  $Z$ 가 어느 실수 구간에 있을 확률은  $CDF(z)$ 를 사용해서 구할 수 있다.



$$P(z_1 \leq Z \leq z_2) = CDF(z_2) - CDF(z_1)$$

### 표준화 (Standardization):

- 확률변수  $X$ 가 정규확률분포를 따르는 경우  $X \sim N(\mu, \sigma^2)$ , 다음의 방식으로  $X$ 를 표준정규확률변수로 변환할 수 있다. 그러면  $Z \sim N(0,1)$ . 이것을 “표준화”라고 부른다.

$$Z = \frac{X - \mu}{\sigma}$$

- 표준화된 값  $x$ 를  $z$ -score “표준점수”라고 부른다.

$$z - score = \frac{x - \mu}{\sigma}$$

- 반대로 표준정규분포 확률변수  $Z$ 를 정규분포  $N(\mu, \sigma^2)$ 를 따르는 확률변수로 변환할 수 있다.

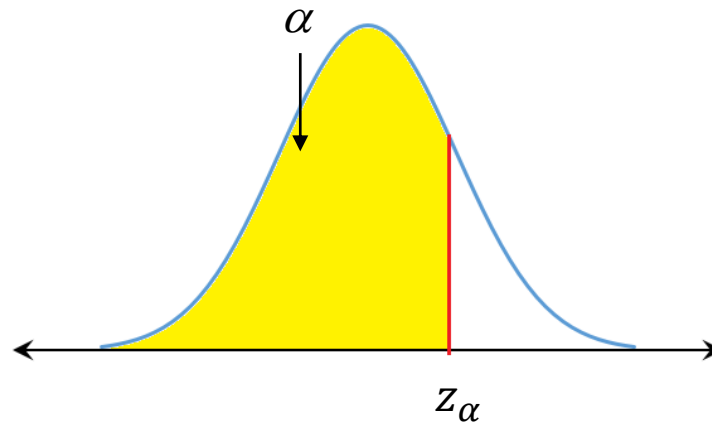
$$X = \sigma Z + \mu$$

## 정규분포 : 표준정규분포의 분위수

### 표준정규분포의 분위수 (Quantile of Standard Normal):

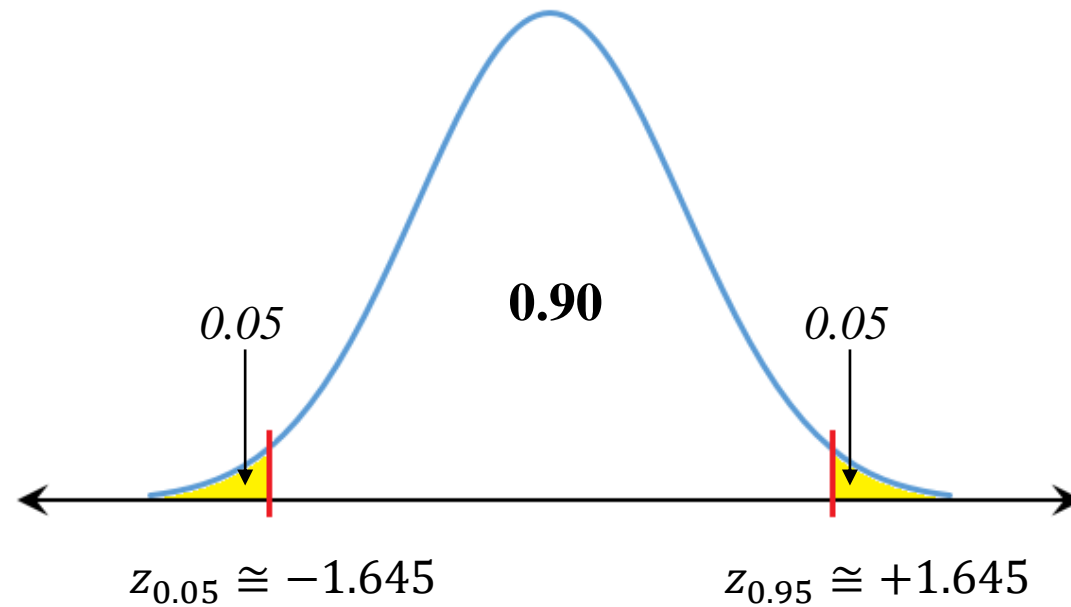
- 분위수 또는 백분위수는 신뢰구간 계산에 필요하다.
- $z_\alpha$ 라고 표기하며 왼쪽 면적(확률 = CDF)이  $\alpha$ 와 같은 위치를 의미한다.

$$P(Z < z_\alpha) = \alpha$$



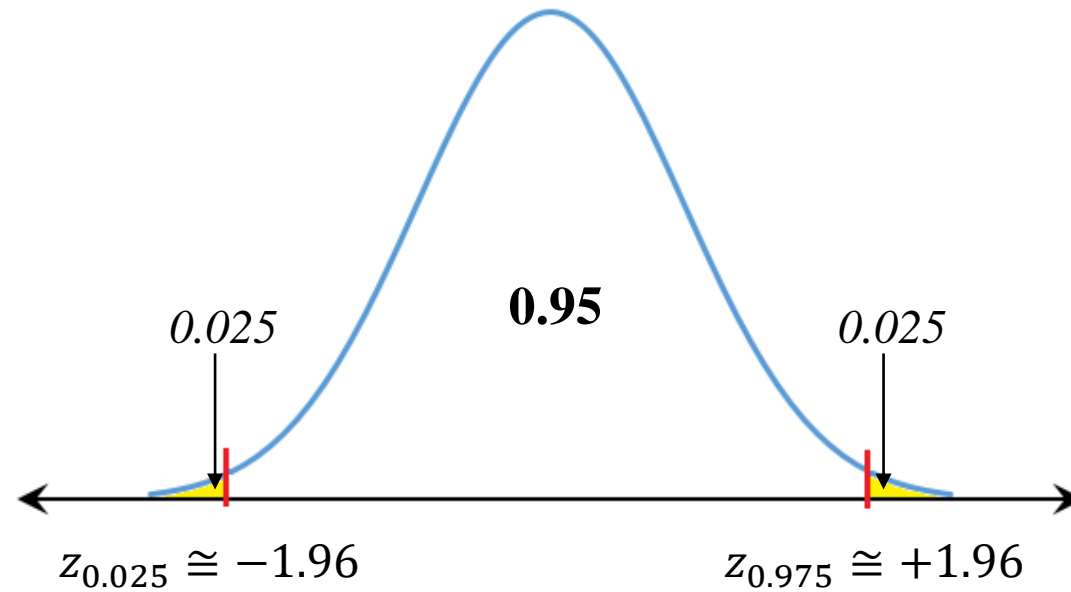
## 정규분포 : 표준정규분포의 분위수

표준정규분포의 분위수 (Quantile of Standard Normal):



## 정규분포 : 표준정규분포의 분위수

표준정규분포의 분위수 (Quantile of Standard Normal):





## 연속확률분포 : 예제 #0203

A군이 시험문제를 푸는데 문항당 평균 50초가 걸리고 표준편차는 20초라고 한다. 48초와 54초 사이에 문항을 풀 확률은 얼마인가? 다음과 같은 **표준정규분포**의 CDF 표를 활용하시오.

<b>Z</b>	<b>CDF(Z)</b>
-0.2	0.4207
-0.1	0.4602
<b>0</b>	<b>0.5</b>
0.1	0.5398
0.2	0.5793

→ 먼저  $x_1 = 48$ 초와  $x_2 = 54$ 초를 표준화 한다.

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{48 - 50}{20} = -\frac{2}{20} = -0.1 \quad z_2 = \frac{x_2 - \mu}{\sigma} = \frac{54 - 50}{20} = \frac{4}{20} = 0.2$$

## 연속확률분포 : 예제 #0203

A군이 시험문제를 푸는데 문항당 평균 50초가 걸리고 표준편차는 20초라고 한다. 48초과 54초 사이에 문항을 풀 확률은 얼마인가? 다음과 같은 **표준정규분포**의 CDF 표를 활용하시오.

<b>Z</b>	<b>CDF(Z)</b>
-0.2	0.4207
-0.1	0.4602
<b>0</b>	<b>0.5</b>
0.1	0.5398
0.2	0.5793

→ CDF를 활용하여 확률을 계산한다.

$$\begin{aligned} P(z_1 \leq Z \leq z_2) &= CDF(z_2) - CDF(z_1) = CDF(0.2) - CDF(-0.1) \\ &= \mathbf{0.5793 - 0.4602 = 0.1191} \end{aligned}$$

### 카이제곱 분포함수 (Chi Square):

- $k$ 개의 **표준**정규분포를 따르는 독립적인 확률변수  $Z \sim N(0,1)$ 가 있을때 카이제곱 확률변수  $Q$ 는 이들의 제곱의 합이다.

$$Q = Z^2 + Z^2 + Z^2 + \cdots + Z^2$$

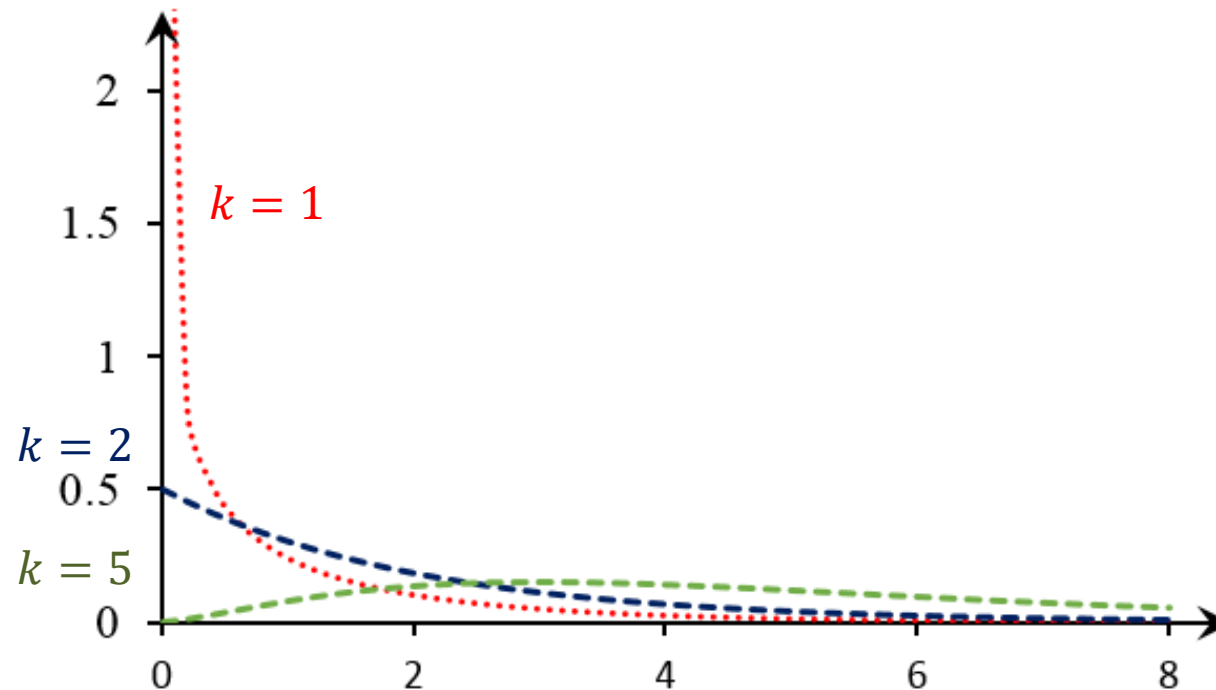
←  $k$  개 →

- 여기에서  $k$ 를 “자유도”라고 부른다.
- 확률변수  $Q$ 가 카이제곱 확률분포를 따른다는 것을 다음과 같이 표기할 수 있다.

$$Q \sim \chi^2(k)$$

## 연속확률분포 : 카이제곱

카이제곱 분포함수 (Chi Square): 자유도  $k$ 의 역할



스튜던트 t 분포함수 (Student t):

- $Q \sim \chi^2(k)$ 이고  $Z \sim N(0,1)$ 일때 스튜던트 t 확률변수  $T$ 는 다음과 같이 정의 된다.

$$T = \frac{Z}{\sqrt{Q/k}}$$

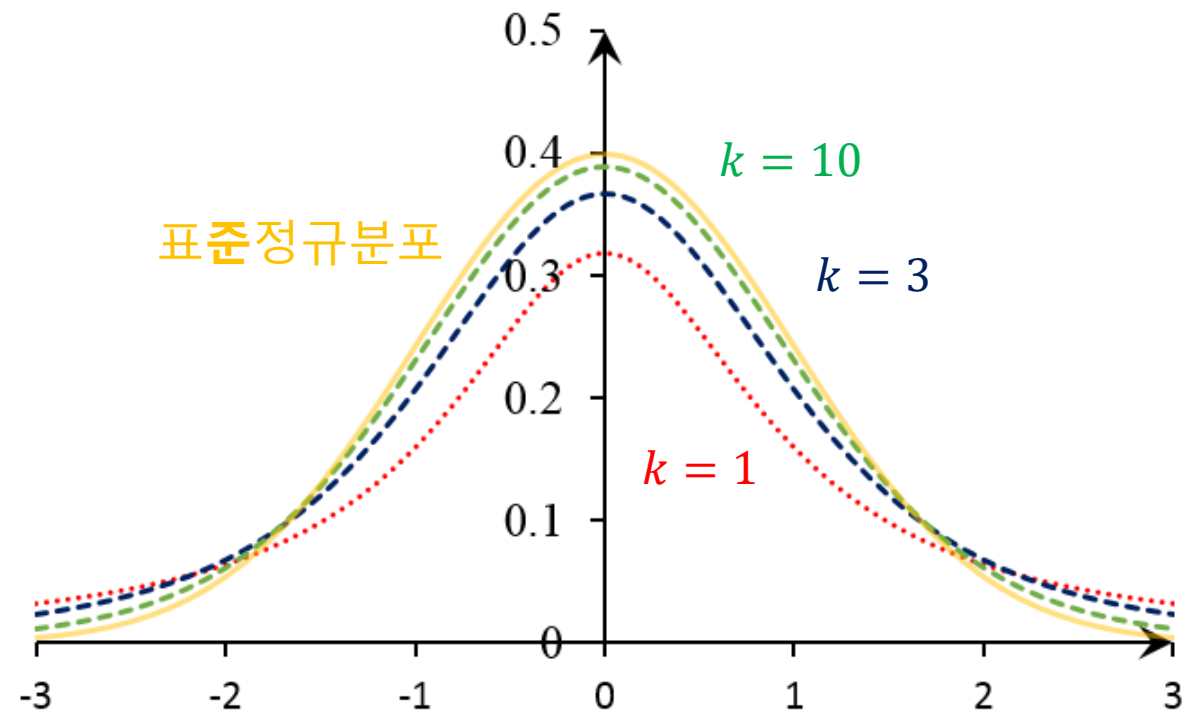
- 여기에서  $k$  는 카이제곱 확률변수의 “자유도”이다.
- 확률변수  $T$ 가 스튜던트 t 확률분포를 따른다는 것을 다음과 같이 표기할 수 있다.

$$T \sim t(k)$$

- 자유도  $k$ 가 커질수록 스튜던트 t는 **표준정규분포**로 수렴한다.

## 연속확률분포 : 스튜던트 $t$

스튜던트  $t$  분포함수 (Student  $t$ ): 자유도  $k$ 의 역할



### F 분포함수:

- $Q_1 \sim \chi^2(d_1)$ 이고  $Q_2 \sim \chi^2(d_2)$ 일때 F 확률변수  $X$ 는 다음과 같이 정의 된다.

$$X = \frac{Q_1/d_1}{Q_2/d_2}$$

- 여기에서  $d_1$ 와  $d_2$ 는 카이제곱 확률변수의 “자유도”이다:

→  $d_1$  = 분자의 자유도

→  $d_2$  = 분모의 자유도

- 확률변수  $X$ 가 F 확률분포를 따른다는 것을 다음과 같이 표기할 수 있다.

$$X \sim F(d_1, d_2)$$

- F 검정, 분산분석 (ANOVA) 등 활용.

## 연속확률분포 : Python 함수

Python의 연속확률 함수:

명칭	함수
연속균등 (Uniform)	<code>scipy.stats.uniform.pdf()</code> ← 확률밀도 <code>scipy.stats.uniform.cdf()</code> ← 누적확률 <code>scipy.stats.uniform.ppf()</code> ← 분위수
정규 (Norm)	<code>scipy.stats.norm.□□□()</code>
지수 (Exponential)	<code>scipy.stats.expon.□□□()</code>
카이제곱 (Chi Square)	<code>scipy.stats.chi2.□□□()</code>
스튜던트 t (Student t)	<code>scipy.stats.t.□□□()</code>
F	<code>scipy.stats.f.□□□()</code>



## 실습 #0202

---

→ 연속확률분포에 대해서 알아보니다. ←

→ 사용: **ex\_0202.ipynb** ←

## 실습 #0203

→ 확률변수를 시뮬레이션 해 봅니다. ←

→ 사용: **ex\_0203.ipynb** ←

## 2. 확률 II:

2.1. 이산확률변수 & 확률분포.

2.2. 이산확률분포의 여러 종류.

2.3. 연속확률변수 & 확률밀도.

2.4. 연속확률밀도의 여러 종류.

2.5. 결합확률과 상관계수.

## 공분산

### 공분산 (Covariance):

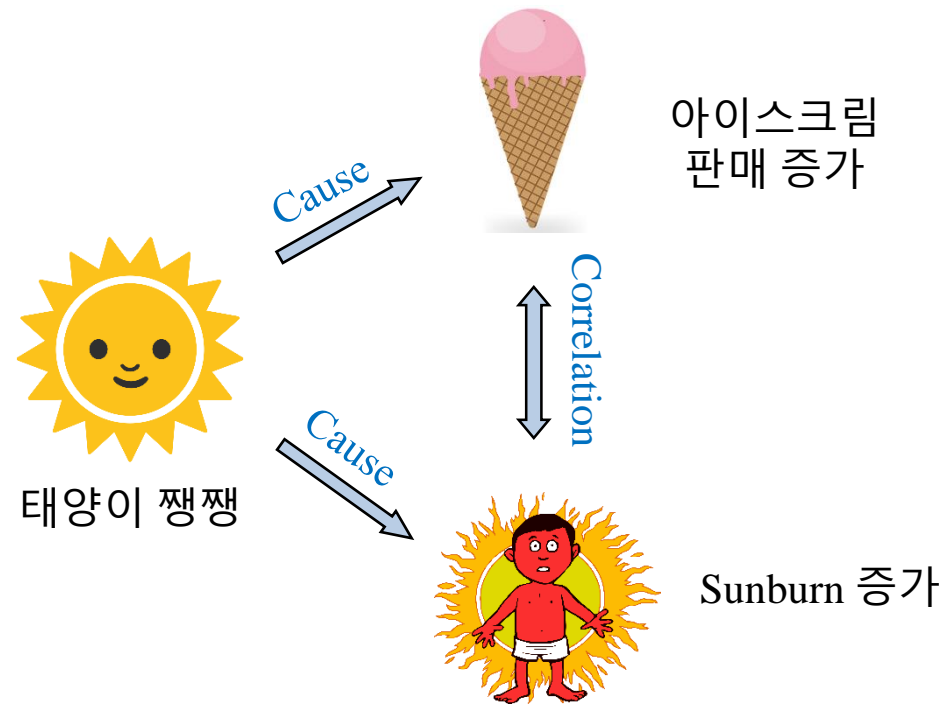
- $Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)] = E[XY] - E[X]E[Y]$        $\Leftarrow$  결합확률을 사용하여 계산.
- $Var(X) = Cov(X, X)$        $\Leftarrow$  “분산과 공분산의 연결”

## 상관계수 (Pearson Correlation Coefficient):

- $Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$
- 상관계수의 값은 -1과 1사이의 수치이다.
- 상관계수는 선형관계의 방향과 강도를 나타낸다.
  - $Corr(X, Y) > 0$  : X와 Y 사이에 **양**의 선형관계가 있음.
  - $Corr(X, Y) < 0$  : X와 Y 사이에 **음**의 선형관계가 있음.
  - $Corr(X, Y) = 0$  : X와 Y 사이에 선형관계가 **없음**.
- 상관성은 원인과 결과로 해석하면 안된다!
- 허구적 상관관계 (spurious correlation)도 있을 수 있으니 주의한다!

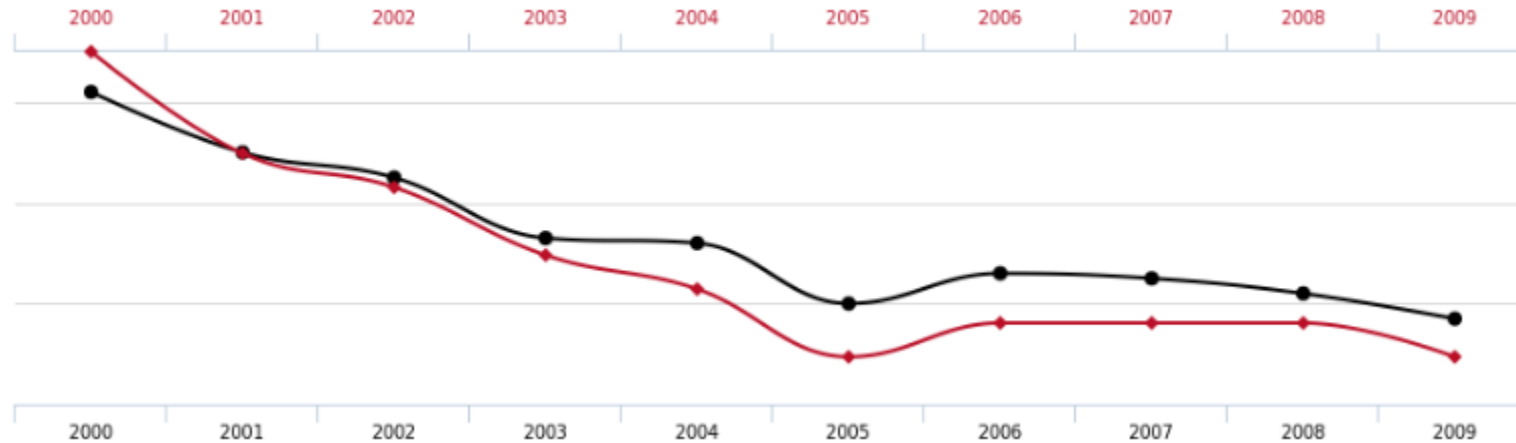
# 상관계수

상관성은 원인과 결과로 해석하면 안된다!



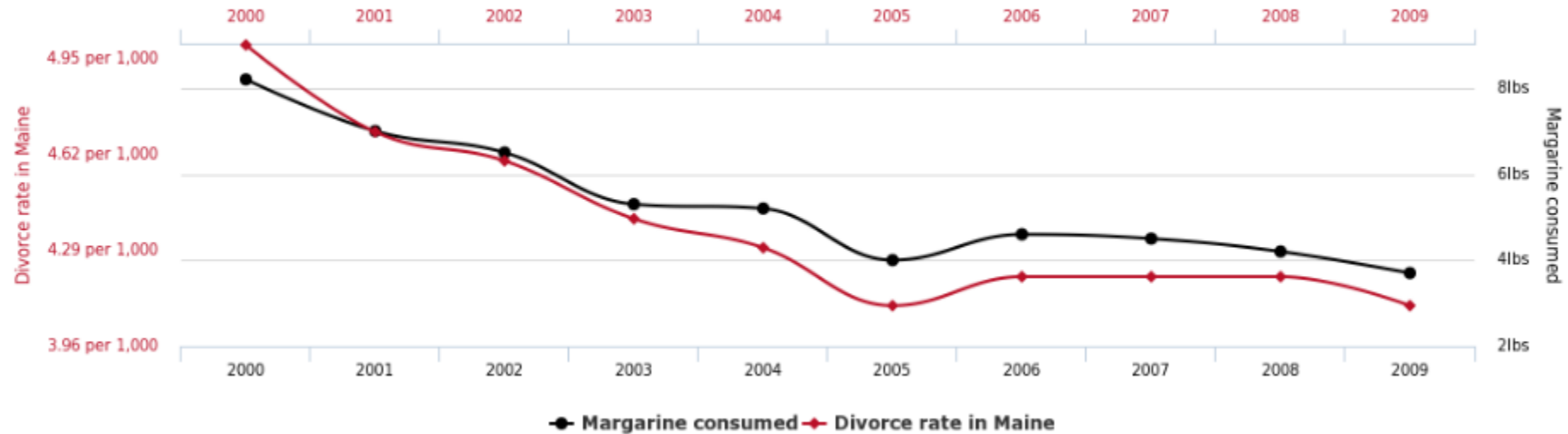
## 상관계수 : 허구적 상관관계

### 허구적 상관관계: 케이스 #1



## 상관계수 : 허구적 상관관계

### 허구적 상관관계: 케이스 #1

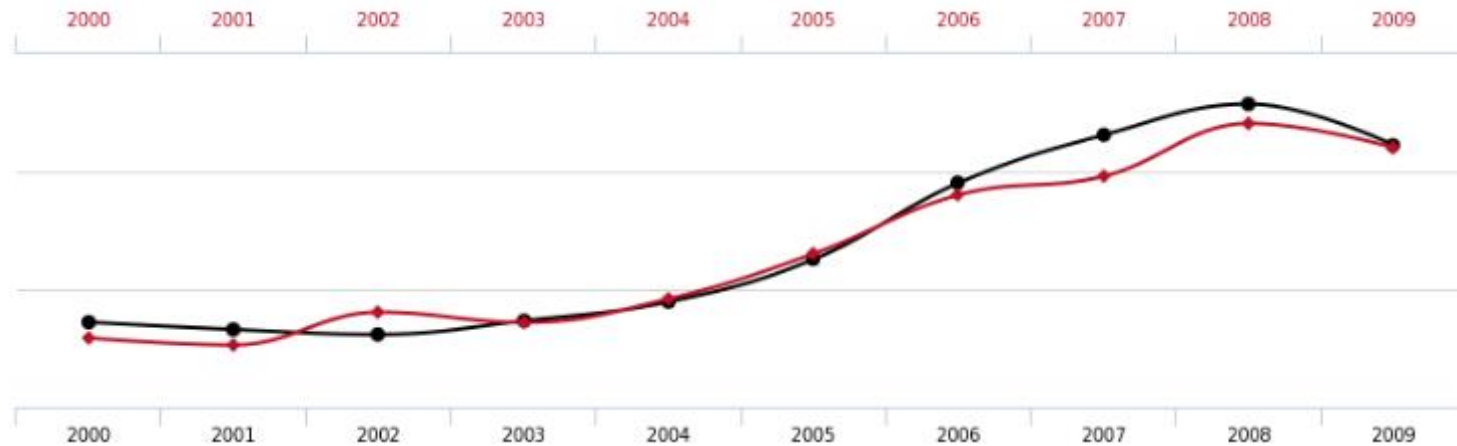


마가린 소비량과 미국 메인(Maine)주의 이혼률



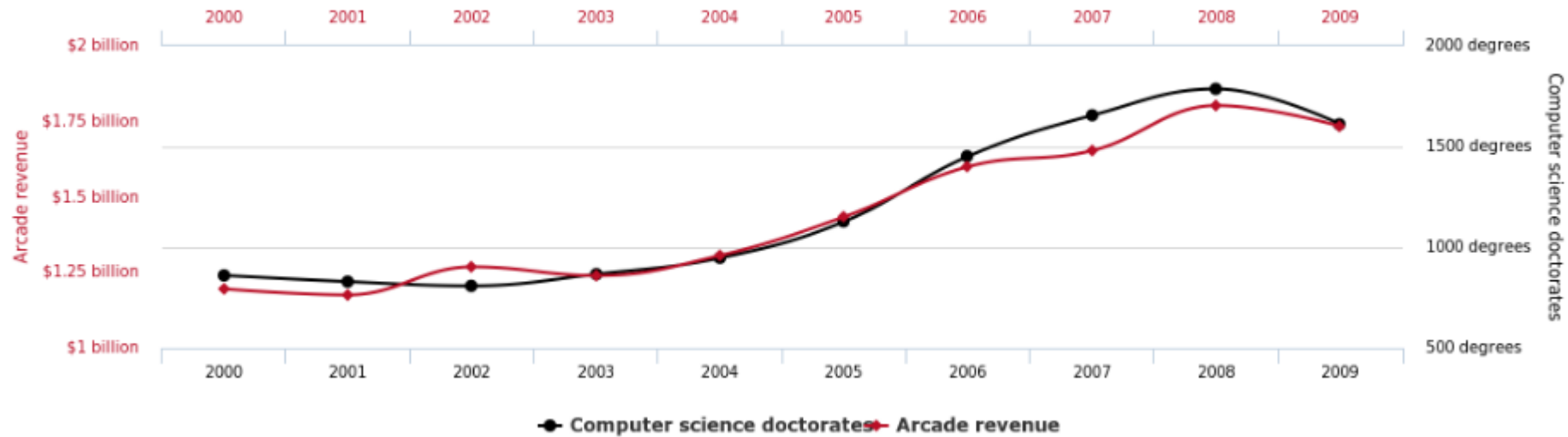
## 상관계수 : 허구적 상관관계

### 허구적 상관관계: 케이스 #2



## 상관계수 : 허구적 상관관계

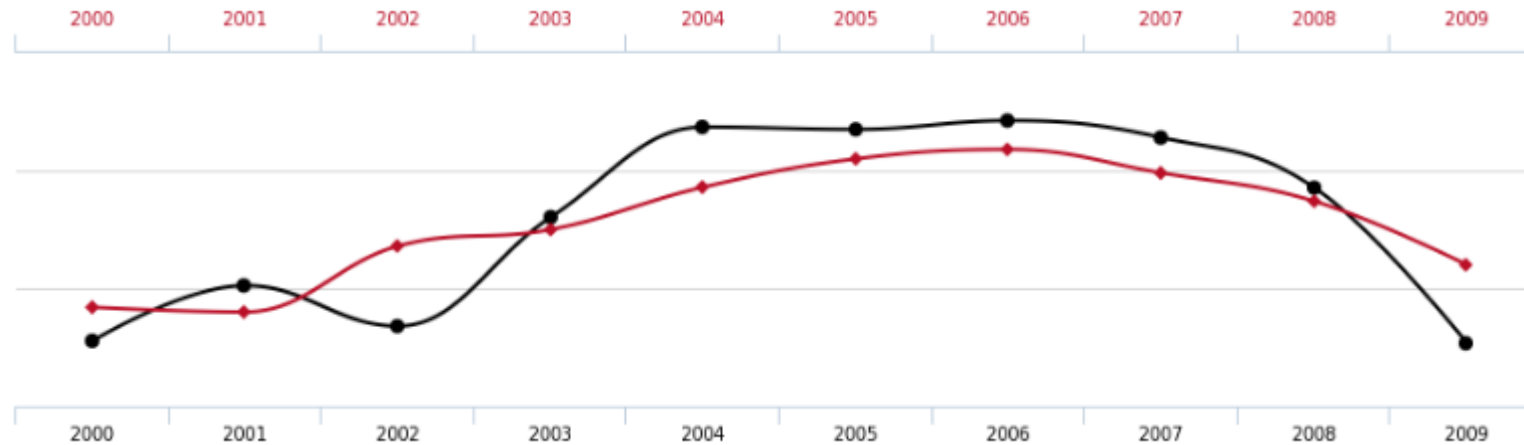
### 허구적 상관관계: 케이스 #2



컴퓨터 공학 박사학위와 아케이드 게임 수익

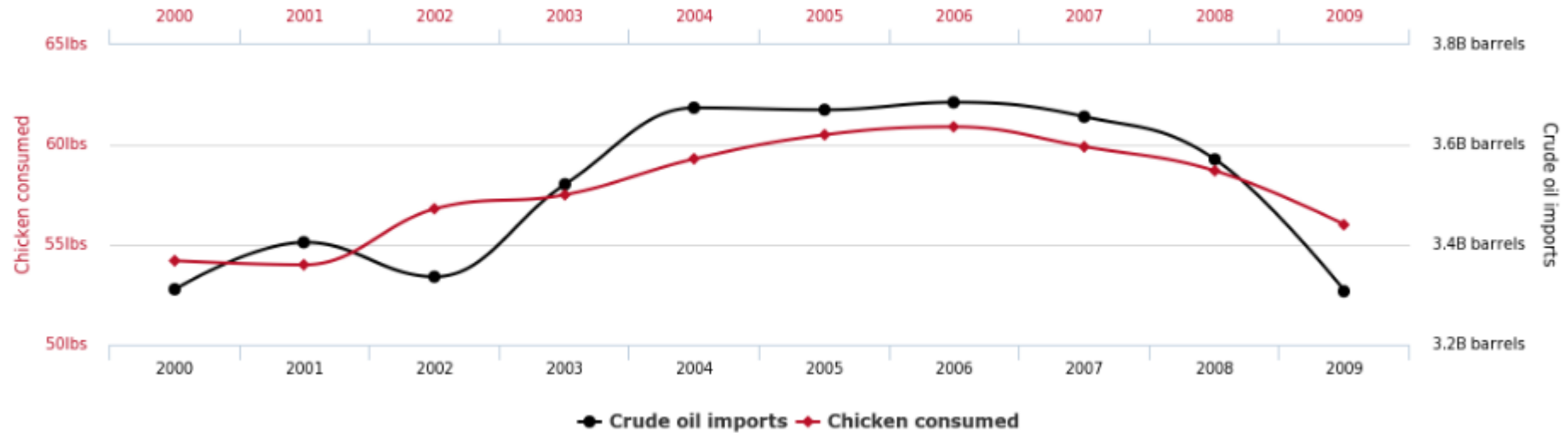
## 상관계수 : 허구적 상관관계

### 허구적 상관관계: 케이스 #3



## 상관계수 : 허구적 상관관계

### 허구적 상관관계: 케이스 #3



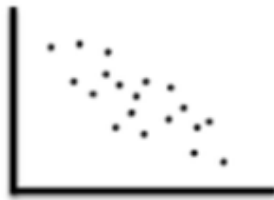
원유 수입과 치킨 소비

## 상관계수

상관계수 (Correlation Coefficient):  $r = \text{Corr}(X, Y)$



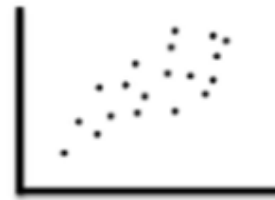
$$r = -1$$



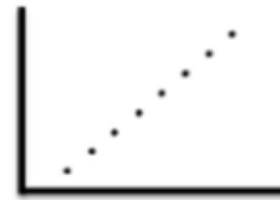
$$-1 < r < 0$$



$$r \approx 0$$



$$0 < r < +1$$

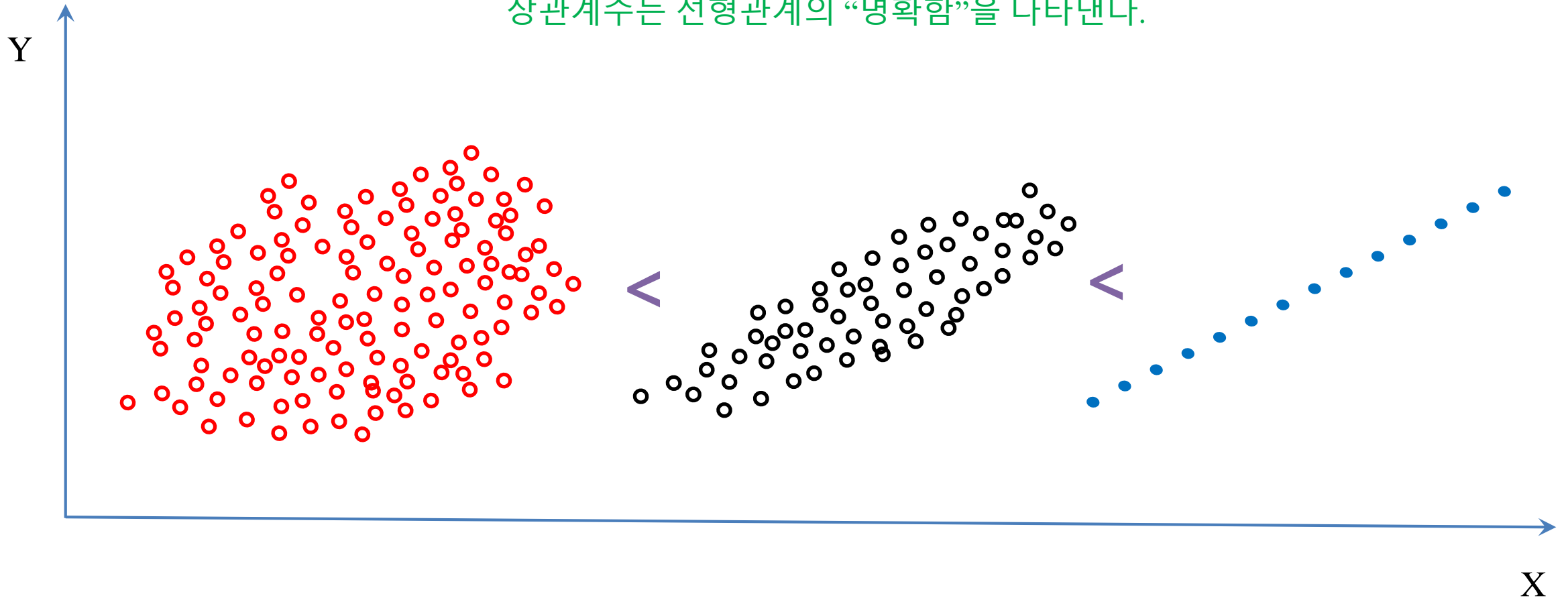


$$r = +1$$

# 상관계수

상관계수 (Correlation Coefficient):

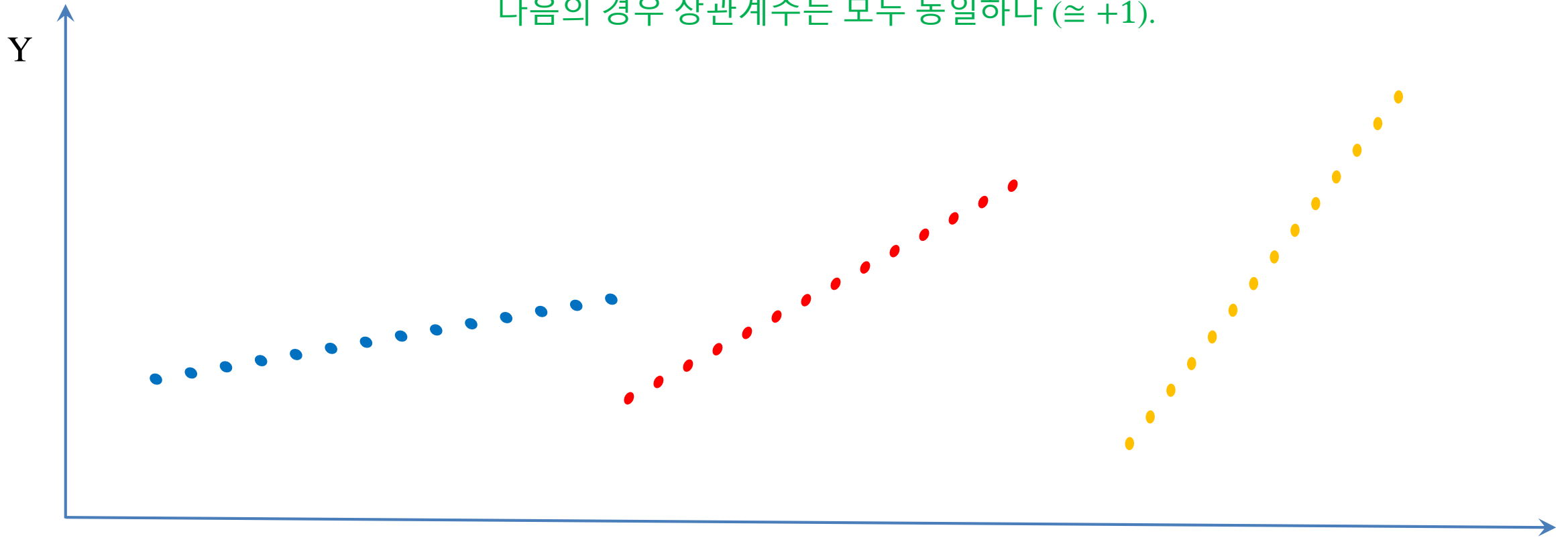
상관계수는 선형관계의 “명확함”을 나타낸다.



# 상관계수

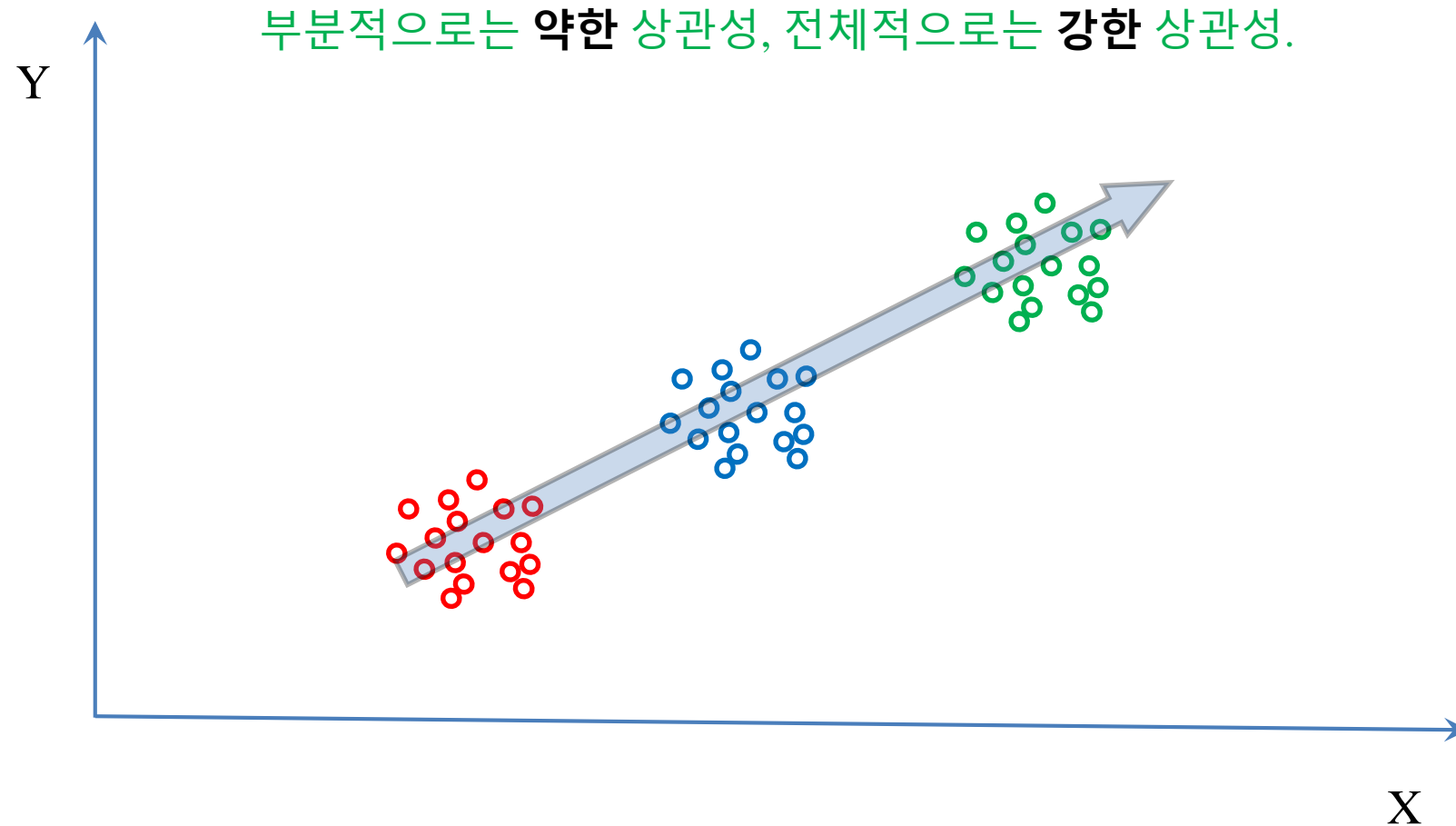
상관계수 (Correlation Coefficient):

다음의 경우 상관계수는 모두 동일하다 ( $\cong +1$ ).



# 상관계수

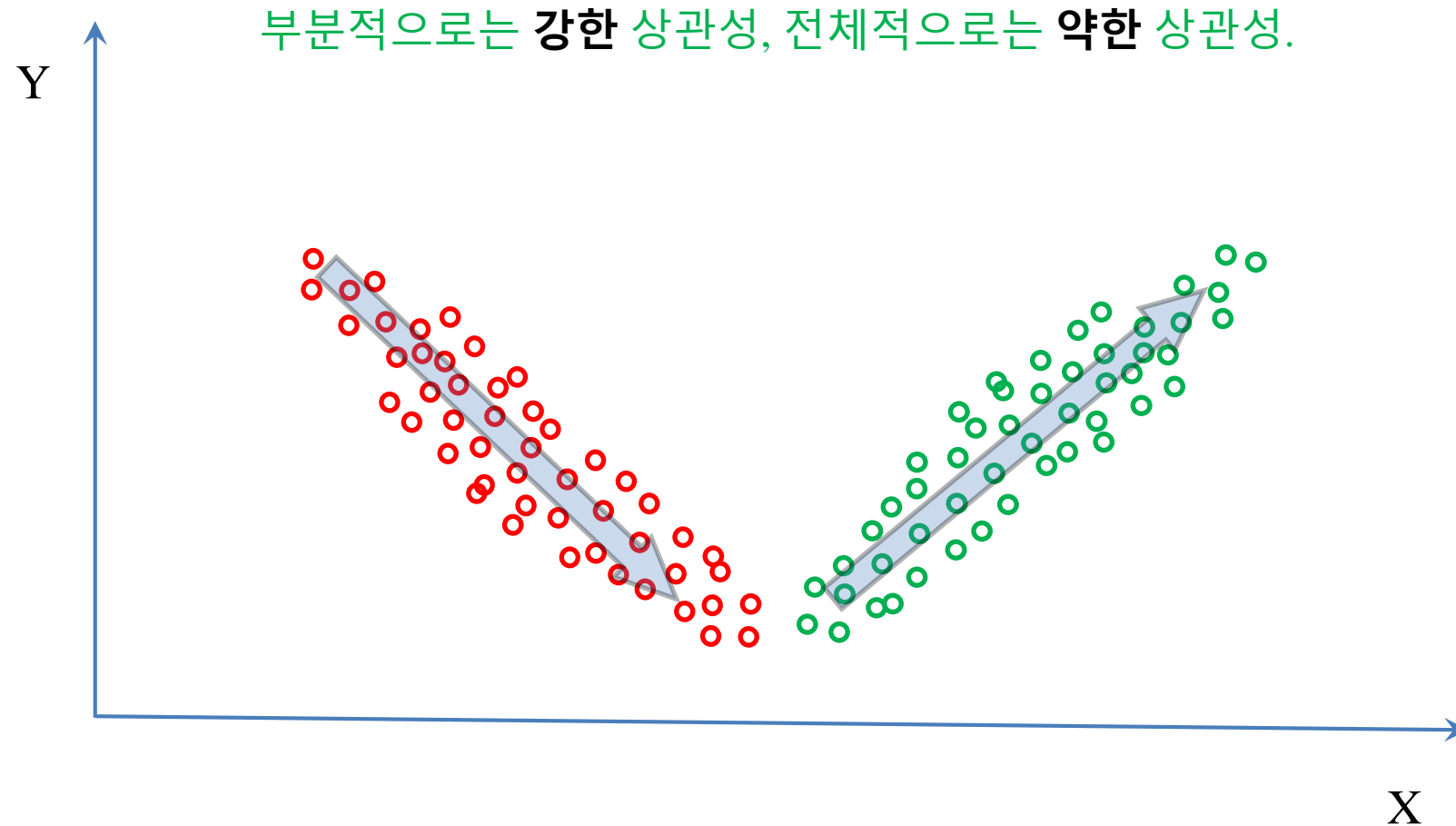
상관계수 (Correlation Coefficient):





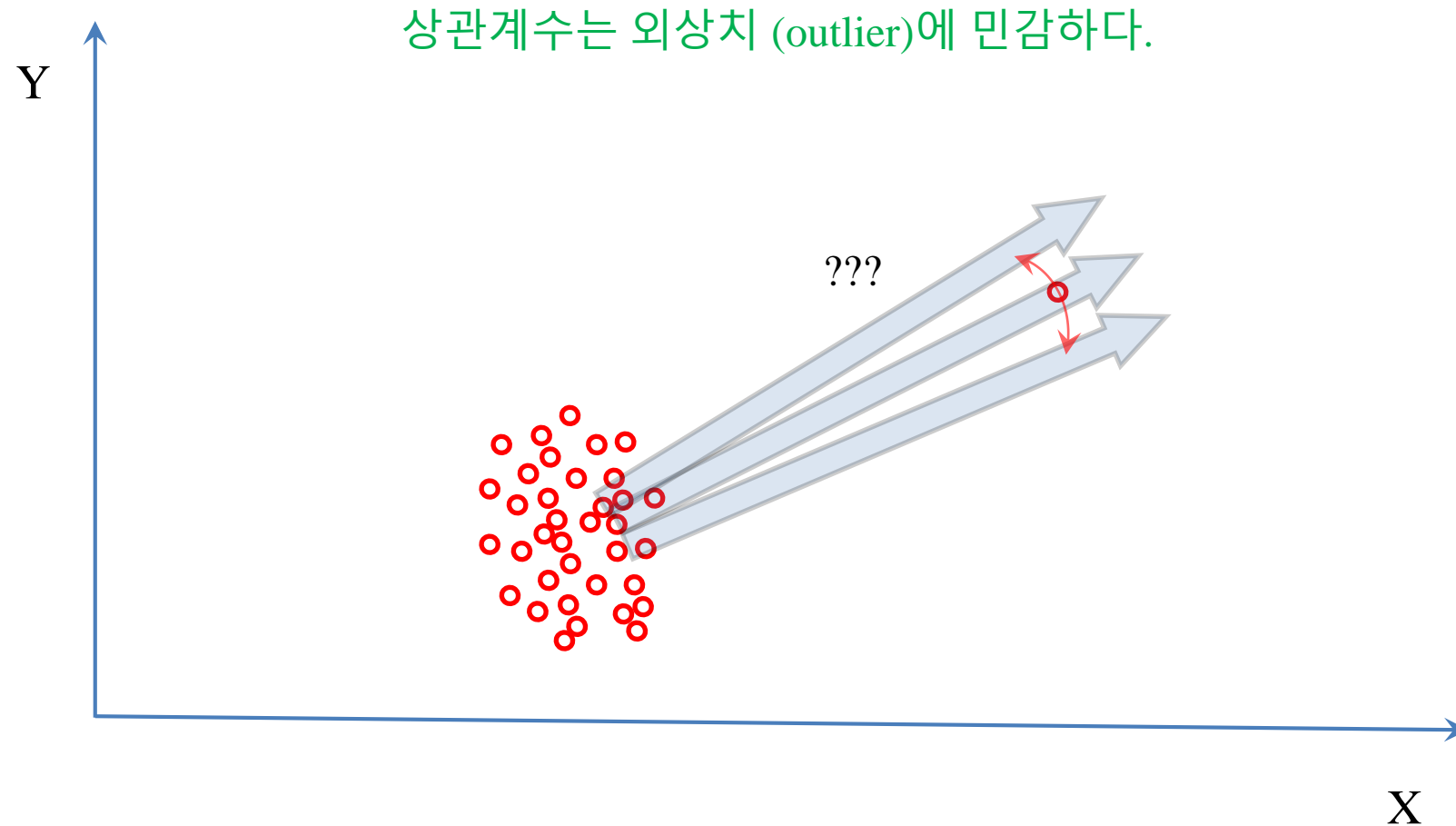
# 상관계수

상관계수 (Correlation Coefficient):



# 상관계수

상관계수 (Correlation Coefficient):



## 독립성 vs 상관성

### 독립성 vs 상관성:

- 독립성:  $P(X, Y) = P(X)P(Y)$ .

$$\rightarrow \text{Cov}(X, Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0.$$

$\rightarrow \text{Corr}(X, Y) = 0$ . 그러므로 “상관성 없음”을 내포함.

- 상관계수:  $\text{Corr}(X, Y)$ .

$\rightarrow$  상관계수는 -1과 1 사이의 수치이다.

$\rightarrow$  “상관성이 없다” = “상관계수 0”. 하지만 독립성을 내포하지는 않는다.

예). -1, 0, 1에서 동일확률을 갖는 확률변수  $X$ 와  $Y = X^2$ 사이의 상관계수는 0

이지만 독립적이지는 않다.

문의:

sychang1@gmail.com